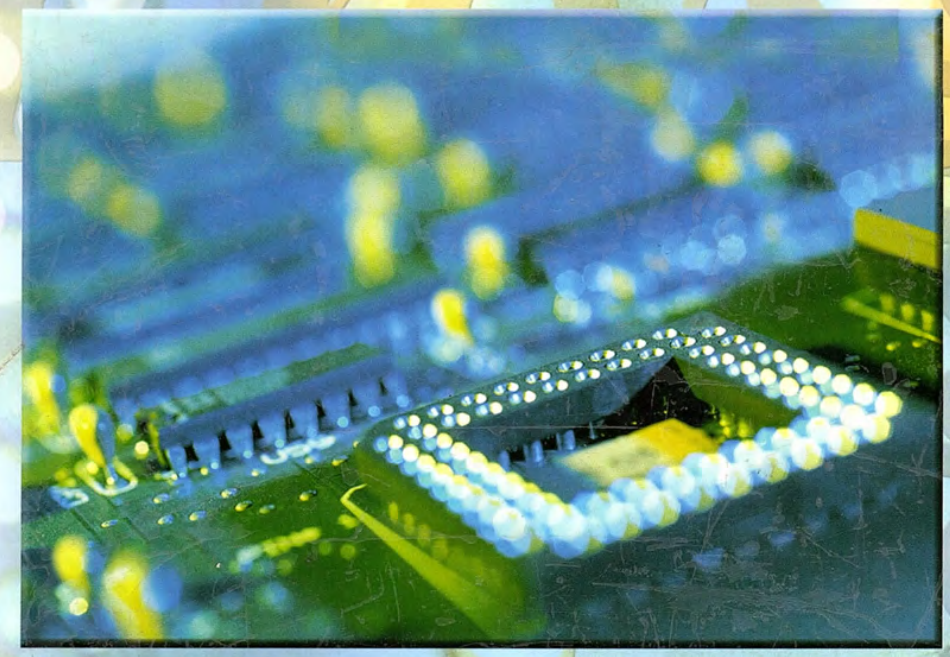

Semiconductor Physics and Devices

Basic Principles

Donald A. Neamen

Third Edition

SAMSUNG EXHIBIT 1026

McGraw-Hill Higher Education

A Division of The McGraw-Hill Companies

SEMICONDUCTOR PHYSICS AND DEVICES: BASIC PRINCIPLES
THIRD EDITION

Published by McGraw-Hill, a business unit of The McGraw-Hill Companies, Inc., 1221 Avenue of the Americas, New York, NY 10020. Copyright © 2003, 1997, 1992 by The McGraw-Hill Companies, Inc. All rights reserved. No part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written consent of The McGraw-Hill Companies, Inc., including, but not limited to, in any network or other electronic storage or transmission, or broadcast for distance learning.

Some ancillaries, including electronic and print components, may not be available to customers outside the United States.

This book is printed on acid-free paper.

International 1 2 3 4 5 6 7 8 9 0 DOC/DOC 0 9 8 7 6 5 4 3 2
Domestic 1 2 3 4 5 6 7 8 9 0 DOC/DOC 0 9 8 7 6 5 4 3 2

ISBN 0–07–232107–5
ISBN 0–07–119862–8 (ISE)

Publisher: *Elizabeth A. Jones*
Senior developmental editor: *Kelley Butcher*
Executive marketing manager: *John Wannemacher*
Project manager: *Joyce Watters*
Production supervisor: *Sherry L. Kane*
Designer: *David W. Hash*
Cover designer: *Rokusek Design*
Cover image: ©*Eyewire, Inc.*
Media project manager: *Sandra M. Schnee*
Media technology senior producer: *Phillip Meek*
Compositor: *Interactive Composition Corporation*
Typeface: *10/12 Times Roman*
Printer: *R. R. Donnelley/Crawfordsville, IN*

Library of Congress Cataloging-in-Publication Data

Neamen, Donald A.
Semiconductor physics and devices : basic principles / Donald A. Neamen. — 3rd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0–07–232107–5 (acid-free paper)
1. Semiconductors. I. Title.

QC611 .N39 2003
537.6'22—dc21 2002019681
CIP

INTERNATIONAL EDITION ISBN 0–07–119862–8
Copyright © 2003. Exclusive rights by The McGraw-Hill Companies, Inc., for manufacture and export. This book cannot be re-exported from the country to which it is sold by McGraw-Hill. The International Edition is not available in North America.

www.mhhe.com

2.1 | PRINCIPLES OF QUANTUM MECHANICS

Before we delve into the mathematics of quantum mechanics, there are three principles we need to consider: the principle of energy quanta, the wave–particle duality principle, and the uncertainty principle.

2.1.1 Energy Quanta

One experiment that demonstrates an inconsistency between experimental results and the classical theory of light is called the photoelectric effect. If monochromatic light is incident on a clean surface of a material, then under certain conditions, electrons (photoelectrons) are emitted from the surface. According to classical physics, if the intensity of the light is large enough, the work function of the material will be overcome and an electron will be emitted from the surface independent of the incident frequency. This result is not observed. The observed effect is that, at a constant incident intensity, the maximum kinetic energy of the photoelectron varies linearly with frequency with a limiting frequency $\nu = \nu_0$, below which no photoelectron is produced. This result is shown in Figure 2.1. If the incident intensity varies at a constant frequency, the rate of photoelectron emission changes, but the maximum kinetic energy remains the same.

Planck postulated in 1900 that thermal radiation is emitted from a heated surface in discrete packets of energy called *quanta*. The energy of these quanta is given by $E = h\nu$, where ν is the frequency of the radiation and h is a constant now known as Planck's constant ($h = 6.625 \times 10^{-34}$ J-s). Then in 1905, Einstein interpreted the photoelectric results by suggesting that the energy in a light wave is also contained in discrete packets or bundles. The particle-like packet of energy is called a *photon,* whose energy is also given by $E = h\nu$. A photon with sufficient energy, then, can knock an electron from the surface of the material. The minimum energy required to remove an electron is called the *work function* of the material

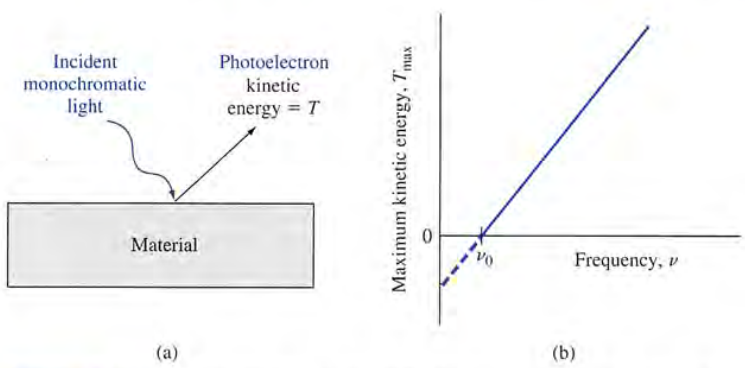

Figure 2.1 | (a) The photoelectric effect and (b) the maximum kinetic energy of the photoelectron as a function of incident frequency.

and any excess photon energy goes into the kinetic energy of the photoelectron. This result was confirmed experimentally as demonstrated in Figure 2.1. The photoelectric effect shows the discrete nature of the photon and demonstrates the particle-like behavior of the photon.

The maximum kinetic energy of the photoelectron can be written as

$$T_{\max} = \frac{1}{2}mv^2 = h\nu - h\nu_0 \quad (\nu \geq \nu_0) \tag{2.1}$$

where $h\nu$ is the incident photon energy and $h\nu_0$ is the minimum energy, or work function, required to remove an electron from the surface.

EXAMPLE 2.1

Objective

To calculate the photon energy corresponding to a particular wavelength.

Consider an x-ray with a wavelength of $\lambda = 0.708 \times 10^{-8}$ cm.

■ Solution

The energy is

$$E = h\nu = \frac{hc}{\lambda} = \frac{(6.625 \times 10^{-34})(3 \times 10^{10})}{0.708 \times 10^{-8}} = 2.81 \times 10^{-15}\ \text{J}$$

This value of energy may be given in the more common unit of electron-volt (see Appendix F). We have

$$E = \frac{2.81 \times 10^{-15}}{1.6 \times 10^{-19}} = 1.75 \times 10^{4}\ \text{eV}$$

■ Comment

The reciprocal relation between photon energy and wavelength is demonstrated: A large energy corresponds to a short wavelength.

2.1.2 Wave–Particle Duality

We have seen in the last section that light waves, in the photoelectric effect, behave as if they are particles. The particle-like behavior of electromagnetic waves was also instrumental in the explanation of the Compton effect. In this experiment, an x-ray beam was incident on a solid. A portion of the x-ray beam was deflected and the frequency of the deflected wave had shifted compared to the incident wave. The observed change in frequency and the deflected angle corresponded exactly to the expected results of a "billiard ball" collision between an x-ray quanta, or photon, and an electron in which both energy and momentum are conserved.

In 1924, de Broglie postulated the existence of matter waves. He suggested that since waves exhibit particle-like behavior, then particles should be expected to show wave-like properties. The hypothesis of de Broglie was the existence of a

We may write the most probable distribution function as

$$\boxed{\frac{N(E)}{g(E)} = f_F(E) = \frac{1}{1 + \exp\left(\dfrac{E - E_F}{kT}\right)}} \tag{3.79}$$

where E_F is called the Fermi energy. The number density $N(E)$ is the number of particles per unit volume per unit energy and the function $g(E)$ is the number of quantum states per unit volume per unit energy. The function $f_F(E)$ is called the Fermi–Dirac distribution or probability function and gives the probability that a quantum state at the energy E will be occupied by an electron. Another interpretation of the distribution function is that $f_F(E)$ is the ratio of filled to total quantum states at any energy E.

3.5.3 The Distribution Function and the Fermi Energy

To begin to understand the meaning of the distribution function and the Fermi energy, we can plot the distribution function versus energy. Initially, let $T = 0$ K and consider the case when $E < E_F$. The exponential term in Equation (3.79) becomes $\exp[(E - E_F)/kT] \to \exp(-\infty) = 0$. The resulting distribution function is $f_F(E < E_F) = 1$. Again let $T = 0$ K and consider the case when $E > E_F$. The exponential term in the distribution function becomes $\exp[(E - E_F)/kT] \to \exp(+\infty) \to +\infty$. The resulting Fermi–Dirac distribution function now becomes $f_F(E > E_F) = 0$.

The Fermi–Dirac distribution function for $T = 0$ K is plotted in Figure 3.27. This result shows that, for $T = 0$ K, the electrons are in their lowest possible energy states. The probability of a quantum state being occupied is unity for $E < E_F$ and the probability of a state being occupied is zero for $E > E_F$. All electrons have energies below the Fermi energy at $T = 0$ K.

Figure 3.28 shows discrete energy levels of a particular system as well as the number of available quantum states at each energy. If we assume, for this case, that

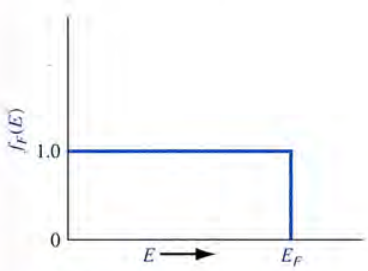

Figure 3.27 | The Fermi probability function versus energy for $T = 0$ K.

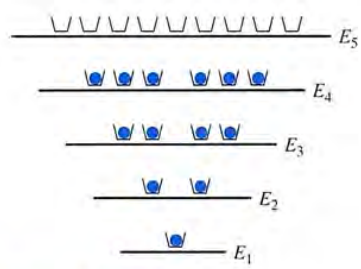

Figure 3.28 | Discrete energy states and quantum states for a particular system at $T = 0$ K.

the system contains 13 electrons, then Figure 3.28 shows how these electrons are distributed among the various quantum states at $T = 0$ K. The electrons will be in the lowest possible energy state, so the probability of a quantum state being occupied in energy levels E_1 through E_4 is unity, and the probability of a quantum state being occupied in energy level E_5 is zero. The Fermi energy, for this case, must be above E_4 but less than E_5. The Fermi energy determines the statistical distribution of electrons and does not have to correspond to an allowed energy level.

Now consider a case in which the density of quantum states $g(E)$ is a continuous function of energy as shown in Figure 3.29. If we have N_0 electrons in this system, then the distribution of these electrons among the quantum states at $T = 0$ K is shown by the dashed line. The electrons are in the lowest possible energy state so that all states below E_F are filled and all states above E_F are empty. If $g(E)$ and N_0 are known for this particular system, then the Fermi energy E_F can be determined.

Consider the situation when the temperature increases above $T = 0$ K. Electrons gain a certain amount of thermal energy so that some electrons can jump to higher energy levels, which means that the distribution of electrons among the available energy states will change. Figure 3.30 shows the same discrete energy levels and quantum states as in Figure 3.28. The distribution of electrons among the quantum states has changed from the $T = 0$ K case. Two electrons from the E_4 level have gained enough energy to jump to E_5, and one electron from E_3 has jumped to E_4. As the temperature changes, the distribution of electrons versus energy changes.

The change in the electron distribution among energy levels for $T > 0$ K can be seen by plotting the Fermi–Dirac distribution function. If we let $E = E_F$ and $T > 0$ K, then Equation (3.79) becomes

$$f_F(E = E_F) = \frac{1}{1 + \exp(0)} = \frac{1}{1+1} = \frac{1}{2}$$

The probability of a state being occupied at $E = E_F$ is $\frac{1}{2}$. Figure 3.31 shows the Fermi–Dirac distribution function plotted for several temperatures, assuming the Fermi energy is independent of temperature.

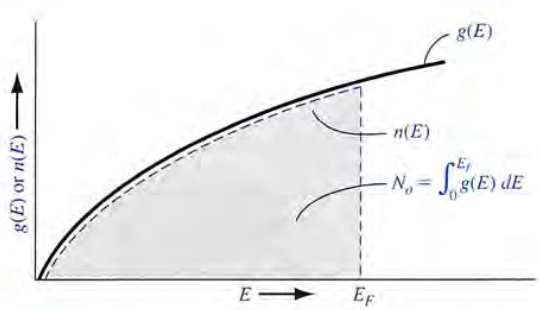

Figure 3.29 | Density of quantum states and electrons in a continuous energy system at $T = 0$ K.

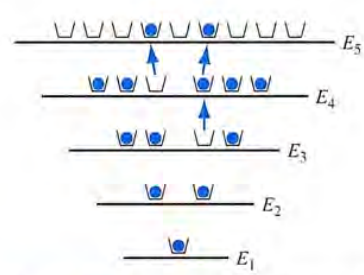

Figure 3.30 | Discrete energy states and quantum states for the same system shown in Figure 3.28 for $T > 0$ K.

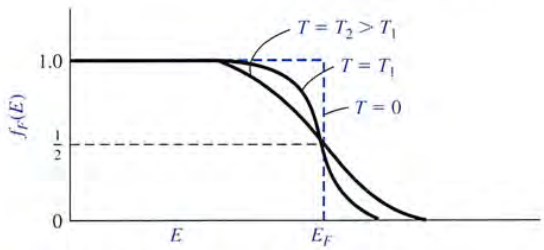

Figure 3.31 | The Fermi probability function versus energy for different temperatures.

We can see that for temperatures above absolute zero, there is a nonzero probability that some energy states above E_F will be occupied by electrons and some energy states below E_F will be empty. This result again means that some electrons have jumped to higher energy levels with increasing thermal energy.

Objective | **EXAMPLE 3.6**

To calculate the probability that an energy state above E_F is occupied by an electron.

Let $T = 300$ K. Determine the probability that an energy level $3kT$ above the Fermi energy is occupied by an electron.

■ Solution

From Equation (3.79), we can write

$$f_F(E) = \frac{1}{1 + \exp\left(\dfrac{E - E_F}{kT}\right)} = \frac{1}{1 + \exp\left(\dfrac{3kT}{kT}\right)}$$

which becomes

$$f_F(E) = \frac{1}{1 + 20.09} = 0.0474 = 4.74\%$$

■ Comment

At energies above E_F, the probability of a state being occupied by an electron can become significantly less than unity, or the ratio of electrons to available quantum states can be quite small.

TEST YOUR UNDERSTANDING

E3.4 Assume the Fermi energy level is 0.30 eV below the conduction band energy.
(*a*) Determine the probability of a state being occupied by an electron at E_c.
(*b*) Repeat part (*a*) for an energy state at $E_c + kT$. Assume $T = 300$ K.
[Ans. (*a*) 9.32×10^{-6}, (*b*) 3.43×10^{-6}]

E3.5 Assume the Fermi energy level is 0.35 eV above the valence band energy. (*a*) Determine the probability of a state being empty of an electron at E_v. (*b*) Repeat part (*a*) for an energy state at $E_V - kT$. Assume $T = 300$ K.
[Ans. (*a*) 1.35×10^{-6}, (*b*) 4.98×10^{-2}]

We can see from Figure 3.31 that the probability of an energy above E_F being occupied increases as the temperature increases and the probability of a state below E_F being empty increases as the temperature increases.

EXAMPLE 3.7

Objective

To determine the temperature at which there is a 1 percent probability that an energy state is empty.

Assume that the Fermi energy level for a particular material is 6.25 eV and that the electrons in this material follow the Fermi–Dirac distribution function. Calculate the temperature at which there is a 1 percent probability that a state 0.30 eV below the Fermi energy level will not contain an electron.

■ Solution

The probability that a state is empty is

$$1 - f_F(E) = 1 - \frac{1}{1 + \exp\left(\dfrac{E - E_F}{kT}\right)}$$

Then

$$0.01 = 1 - \frac{1}{1 + \exp\left(\dfrac{5.95 - 6.25}{kT}\right)}$$

Solving for kT, we find $kT = 0.06529$ eV, so that the temperature is $T = 756$ K.

■ Comment

The Fermi probability function is a strong function of temperature.

TEST YOUR UNDERSTANDING

E3.6 Repeat Exercise E3.4 for $T = 400$ K. [Ans. (*a*) 9.20×10^{-5}, (*b*) 1.69×10^{-4}]
E3.7 Repeat Exercise E3.5 for $T = 400$ K. [Ans. (*a*) 1.46×10^{-5}, (*b*) 3.96×10^{-5}]

We may note that the probability of a state a distance dE above E_F being occupied is the same as the probability of a state a distance dE below E_F being empty. The function $f_F(E)$ is symmetrical with the function $1 - f_F(E)$ about the Fermi energy, E_F. This symmetry effect is shown in Figure 3.32 and will be used in the next chapter.

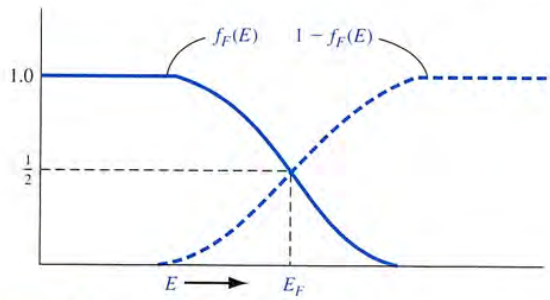

Figure 3.32 | The probability of a state being occupied, $f_F(E)$, and the probability of a state being empty, $1 - f_F(E)$.

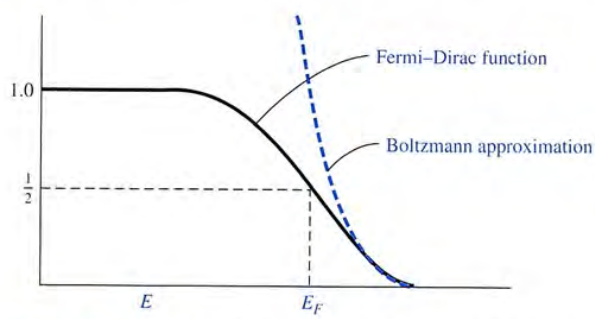

Figure 3.33 | The Fermi–Dirac probability function and the Maxwell–Boltzmann approximation.

Consider the case when $E - E_F \gg kT$, where the exponential term in the denominator of Equation (3.79) is much greater than unity. We may neglect the 1 in the denominator, so the Fermi–Dirac distribution function becomes

$$\boxed{f_F(E) \approx \exp\left[\frac{-(E - E_F)}{kT}\right]} \tag{3.80}$$

Equation (3.80) is known as the Maxwell–Boltzmann approximation, or simply the Boltzmann approximation, to the Fermi–Dirac distribution function. Figure 3.33 shows the Fermi–Dirac probability function and the Boltzmann approximation. This figure gives an indication of the range of energies over which the approximation is valid.

Objective | **EXAMPLE 3.8**

To determine the energy at which the Boltzmann approximation may be considered valid.

Calculate the energy, in terms of kT and E_F, at which the difference between the Boltzmann approximation and the Fermi–Dirac function is 5 percent of the Fermi function.

■ Solution

We can write

$$\frac{\exp\left[\dfrac{-(E-E_F)}{kT}\right] - \dfrac{1}{1+\exp\left(\dfrac{E-E_F}{kT}\right)}}{\dfrac{1}{1+\exp\left(\dfrac{E-E_F}{kT}\right)}} = 0.05$$

If we multiply both numerator and denominator by the $1 + \exp(\ )$ function, we have

$$\exp\left[\frac{-(E-E_F)}{kT}\right] \cdot \left\{1 + \exp\left[\frac{E-E_F}{kT}\right]\right\} - 1 = 0.05$$

which becomes

$$\exp\left[\frac{-(E-E_F)}{kT}\right] = 0.05$$

or

$$(E - E_F) = kT\, ln\left(\frac{1}{0.05}\right) \approx 3kT$$

■ Comment

As seen in this example and in Figure 3.33, the $E - E_F \gg kT$ notation is somewhat misleading. The Maxwell–Boltzmann and Fermi–Dirac functions are within 5 percent of each other when $E - E_F \approx 3kT$.

The actual Boltzmann approximation is valid when $\exp[(E - E_F)/kT] \gg 1$. However, it is still common practice to use the $E - E_F \gg kT$ notation when applying the Boltzmann approximation. We will use this Boltzmann approximation in our discussion of semiconductors in the next chapter.

3.6 | SUMMARY

- Discrete allowed electron energies split into a band of allowed energies as atoms are brought together to form a crystal.
- The concept of allowed and forbidden energy bands was developed more rigorously by considering quantum mechanics and Schrodinger's wave equation using the Kronig–Penney model representing the potential function of a single crystal material. This result forms the basis of the energy band theory of semiconductors.
- The concept of effective mass was developed. Effective mass relates the motion of a particle in a crystal to an externally applied force and takes into account the effect of the crystal lattice on the motion of the particle.
- Two charged particles exist in a semiconductor. An electron is a negatively charged particle with a positive effective mass existing at the bottom of an allowed energy band. A hole is a positively charged particle with a positive effective mass existing at the top of an allowed energy band.

- The E versus k diagram of silicon and gallium arsenide were given and the concept of direct and indirect bandgap semiconductors was discussed.
- Energies within an allowed energy band are actually at discrete levels and each contains a finite number of quantum states. The density per unit energy of quantum states was determined by using the three-dimensional infinite potential well as a model.
- In dealing with large numbers of electrons and holes, we must consider the statistical behavior of these particles. The Fermi–Dirac probability function was developed, which gives the probability of a quantum state at an energy E of being occupied by an electron. The Fermi energy was defined.

GLOSSARY OF IMPORTANT TERMS

allowed energy band A band or range of energy levels that an electron in a crystal is allowed to occupy based on quantum mechanics.

density of states function The density of available quantum states as a function of energy, given in units of number per unit energy per unit volume.

electron effective mass The parameter that relates the acceleration of an electron in the conduction band of a crystal to an external force; a parameter that takes into account the effect of internal forces in the crystal.

Fermi–Dirac probability function The function describing the statistical distribution of electrons among available energy states and the probability that an allowed energy state is occupied by an electron.

fermi energy In the simplest definition, the energy below which all states are filled with electrons and above which all states are empty at $T = 0$ K.

forbidden energy band A band or range of energy levels that an electron in a crystal is not allowed to occupy based on quantum mechanics.

hole The positively charged "particle" associated with an empty state in the top of the valence band.

hole effective mass The parameter that relates the acceleration of a hole in the valence band of a crystal to an applied external force (a positive quantity); a parameter that takes into account the effect of internal forces in a crystal.

k-space diagram The plot of electron energy in a crystal versus k, where k is the momentum-related constant of the motion that incorporates the crystal interaction.

Kronig–Penney model The mathematical model of a periodic potential function representing a one-dimensional single-crystal lattice by a series of periodic step functions.

Maxwell–Boltzmann approximation The condition in which the energy is several kT above the Fermi energy or several kT below the Fermi energy so that the Fermi–Dirac probability function can be approximated by a simple exponential function.

Pauli exclusion principle The principle which states that no two electrons can occupy the same quantum state.

CHECKPOINT

After studying this chapter, the reader should have the ability to:

- Discuss the concept of allowed and forbidden energy bands in a single crystal both qualitatively and more rigorously from the results of using the Kronig–Penney model.

CHAPTER 4

The Semiconductor in Equilibrium

PREVIEW

So far, we have been considering a general crystal and applying to it the concepts of quantum mechanics in order to determine a few of the characteristics of electrons in a single-crystal lattice. In this chapter, we will apply these concepts specifically to a semiconductor material. In particular, we will use the density of quantum states in the conduction band and the density of quantum states in the valence band along with the Fermi–Dirac probability function to determine the concentration of electrons and holes in the conduction and valence bands, respectively. We will also apply the concept of the Fermi energy to the semiconductor material.

This chapter deals with the semiconductor in equilibrium. Equilibrium, or thermal equilibrium, implies that no external forces such as voltages, electric fields, magnetic fields, or temperature gradients are acting on the semiconductor. All properties of the semiconductor will be independent of time in this case. Equilibrium is our starting point for developing the physics of the semiconductor. We will then be able to determine the characteristics that result when deviations from equilibrium occur, such as when a voltage is applied to a semiconductor device.

We will initially consider the properties of an intrinsic semiconductor, that is, a pure crystal with no impurity atoms or defects. We will see that the electrical properties of a semiconductor can be altered in desirable ways by adding controlled amounts of specific impurity atoms, called *dopant atoms,* to the crystal. Depending upon the type of dopant atom added, the dominant charge carrier in the semiconductor will be either electrons in the conduction band or holes in the valence band. Adding dopant atoms changes the distribution of electrons among the available energy states, so the Fermi energy becomes a function of the type and concentration of impurity atoms.

Finally, as part of this discussion, we will attempt to add more insight into the significance of the Fermi energy. ■

4.1 | CHARGE CARRIERS IN SEMICONDUCTORS

Current is the rate at which charge flows. In a semiconductor, two types of charge carrier, the electron and the hole, can contribute to a current. Since the current in a semiconductor is determined largely by the number of electrons in the conduction band and the number of holes in the valence band, an important characteristic of the semiconductor is the density of these charge carriers. The density of electrons and holes is related to the density of states function and the Fermi distribution function, both of which we have considered. A qualitative discussion of these relationships will be followed by a more rigorous mathematical derivation of the thermal-equilibrium concentration of electrons and holes.

4.1.1 Equilibrium Distribution of Electrons and Holes

The distribution (with respect to energy) of electrons in the conduction band is given by the density of allowed quantum states times the probability that a state is occupied by an electron. This statement is written in equation form as

$$n(E) = g_c(E) f_F(E) \tag{4.1}$$

where $f_F(E)$ is the Fermi–Dirac probability function and $g_c(E)$ is the density of quantum states in the conduction band. The total electron concentration per unit volume in the conduction band is then found by integrating Equation (4.1) over the entire conduction-band energy.

Similarly, the distribution (with respect to energy) of holes in the valence band is the density of allowed quantum states in the valence band multiplied by the probability that a state is *not* occupied by an electron. We may express this as

$$p(E) = g_v(E)[1 - f_F(E)] \tag{4.2}$$

The total hole concentration per unit volume is found by integrating this function over the entire valence-band energy.

To find the thermal-equilibrium electron and hole concentrations, we need to determine the position of the Fermi energy E_F with respect to the bottom of the conduction-band energy E_c and the top of the valence-band energy E_v. To address this question, we will initially consider an intrinsic semiconductor. An ideal intrinsic semiconductor is a pure semiconductor with no impurity atoms and no lattice defects in the crystal (e.g., pure silicon). We have argued in the previous chapter that, for an intrinsic semiconductor at $T = 0$ K, all energy states in the valence band are filled with electrons and all energy states in the conduction band are empty of electrons. The Fermi energy must, therefore, be somewhere between E_c and E_v. (The Fermi energy does not need to correspond to an allowed energy.)

As the temperature begins to increase above 0 K, the valence electrons will gain thermal energy. A few electrons in the valence band may gain sufficient energy to jump to the conduction band. As an electron jumps from the valence band to the conduction band, an empty state, or hole, is created in the valence band. In an intrinsic semiconductor, then, electrons and holes are created in pairs by the thermal energy so

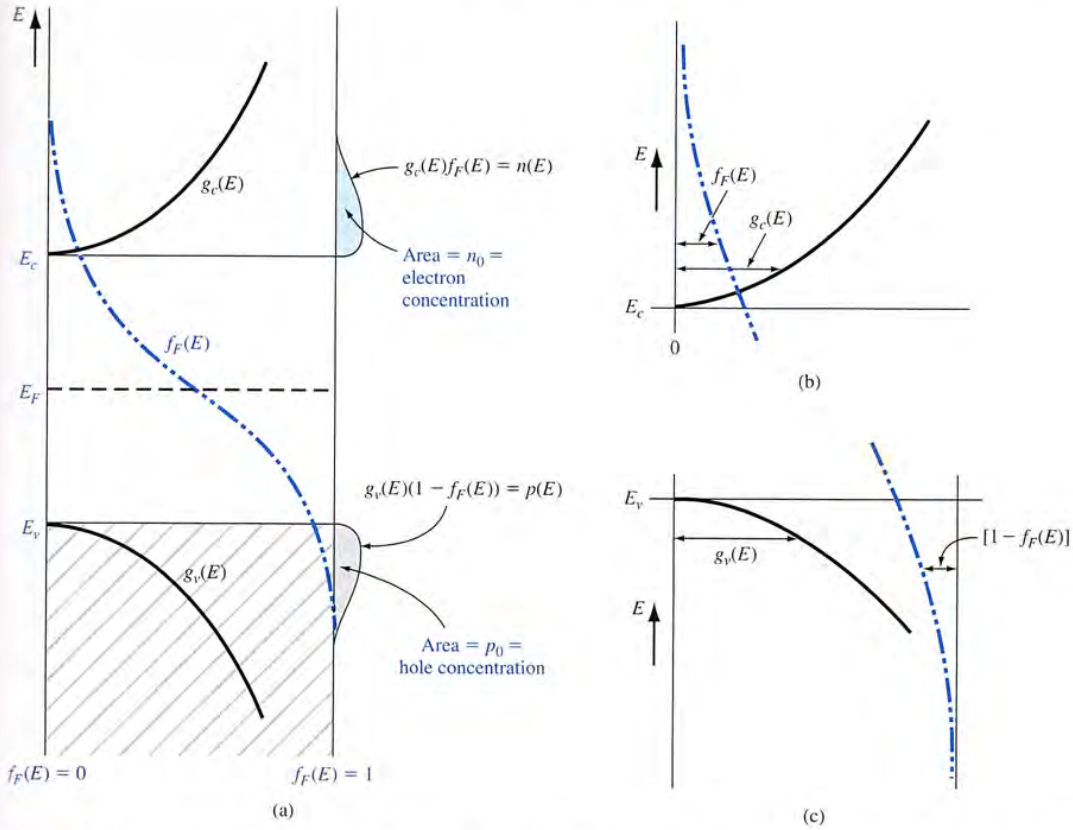

Figure 4.1 | (a) Density of states functions, Fermi–Dirac probability function, and areas representing electron and hole concentrations for the case when E_F is near the midgap energy; (b) expanded view near the conduction band energy; and (c) expanded view near the valence band energy.

that the number of electrons in the conduction band is equal to the number of holes in the valence band.

Figure 4.1a shows a plot of the density of states function in the conduction band $g_c(E)$, the density of states function in the valence band $g_v(E)$, and the Fermi–Dirac probability function for $T > 0$ K when E_F is approximately halfway between E_c and E_v. If we assume, for the moment, that the electron and hole effective masses are equal, then $g_c(E)$ and $g_v(E)$ are symmetrical functions about the midgap energy (the energy midway between E_c and E_v). We noted previously that the function $f_F(E)$ for $E > E_F$ is symmetrical to the function $1 - f_F(E)$ for $E < E_F$ about the energy $E = E_F$. This also means that the function $f_F(E)$ for $E = E_F + dE$ is equal to the function $1 - f_F(E)$ for $E = E_F - dE$.

or

$$E_{Fi} - E_{\text{midgap}} = -0.0128 \text{ eV} = -12.8 \text{ meV}$$

■ Comment

The intrinsic Fermi level in silicon is 12.8 meV below the midgap energy. If we compare 12.8 meV to 560 meV, which is one-half of the bandgap energy of silicon, we can, in many applications, simply approximate the intrinsic Fermi level to be in the center of the bandgap.

TEST YOUR UNDERSTANDING

E4.6 Determine the position of the intrinsic Fermi level with respect to the center of the bandgap in GaAs at $T = 300$ K. (Ans. -38.2 meV)

4.2 | DOPANT ATOMS AND ENERGY LEVELS

The intrinsic semiconductor may be an interesting material, but the real power of semiconductors is realized by adding small, controlled amounts of specific dopant, or impurity, atoms. This doping process, described briefly in Chapter 1, can greatly alter the electrical characteristics of the semiconductor. The doped semiconductor, called an *extrinsic* material, is the primary reason we can fabricate the various semiconductor devices that we will consider in later chapters.

4.2.1 Qualitative Description

In Chapter 3, we discussed the covalent bonding of silicon and considered the simple two-dimensional representation of the single-crystal silicon lattice as shown in Figure 4.3. Now consider adding a group V element, such as phosphorus, as a substitutional impurity. The group V element has five valence electrons. Four of these will contribute to the covalent bonding with the silicon atoms, leaving the fifth more loosely bound to the phosphorus atom. This effect is schematically shown in Figure 4.4. We refer to the fifth valence electron as a donor electron.

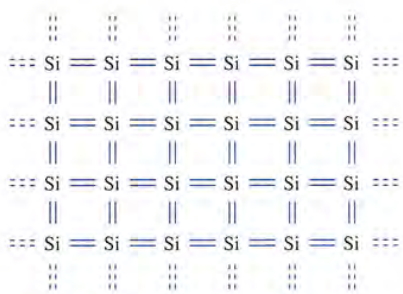

Figure 4.3 | Two-dimensional representation of the intrinsic silicon lattice.

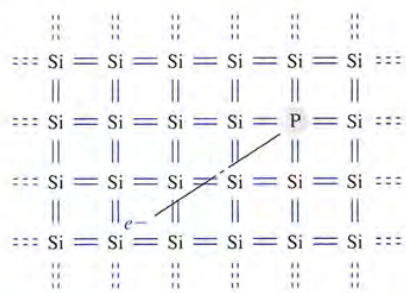

Figure 4.4 | Two-dimensional representation of the silicon lattice doped with a phosphorus atom.

The phosphorus atom without the donor electron is positively charged. At very low temperatures, the donor electron is bound to the phosphorus atom. However, by intuition, it should seem clear that the energy required to elevate the donor electron into the conduction band is considerably less than that for the electrons involved in the covalent bonding. Figure 4.5 shows the energy-band diagram that we would expect. The energy level, E_d, is the energy state of the donor electron.

If a small amount of energy, such as thermal energy, is added to the donor electron, it can be elevated into the conduction band, leaving behind a positively charged phosphorus ion. The electron in the conduction band can now move through the crystal generating a current, while the positively charged ion is fixed in the crystal. This type of impurity atom donates an electron to the conduction band and so is called a *donor impurity atom.* The donor impurity atoms add electrons to the conduction band without creating holes in the valence band. The resulting material is referred to as an *n-type* semiconductor (*n* for the negatively charged electron).

Now consider adding a group III element, such as boron, as a substitutional impurity to silicon. The group III element has three valence electrons, which are all taken up in the covalent bonding. As shown in Figure 4.6a, one covalent bonding position appears to be empty. If an electron were to occupy this "empty" position, its

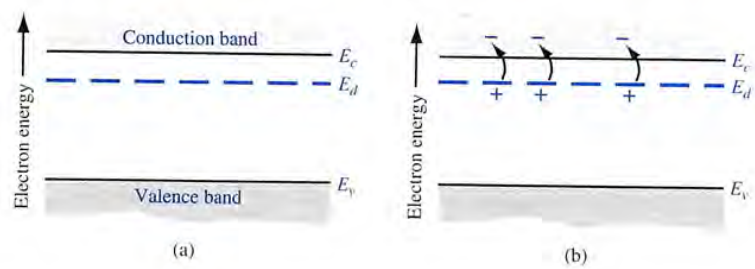

Figure 4.5 | The energy-band diagram showing (a) the discrete donor energy state and (b) the effect of a donor state being ionized.

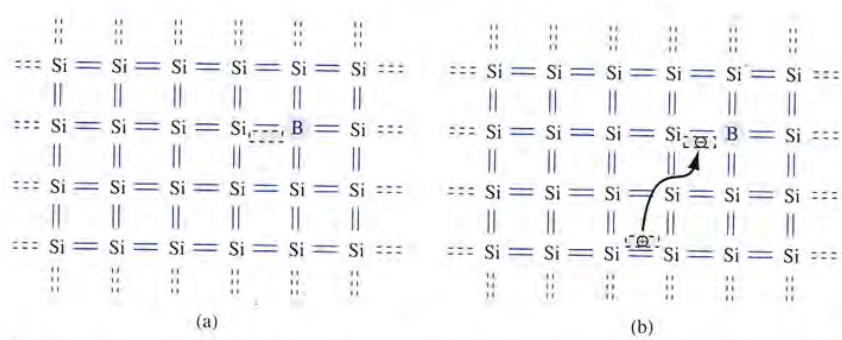

Figure 4.6 | Two-dimensional representation of a silicon lattice (a) doped with a boron atom and (b) showing the ionization of the boron atom resulting in a hole.

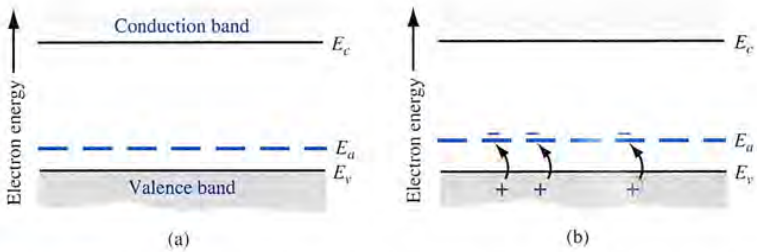

Figure 4.7 | The energy-band diagram showing (a) the discrete acceptor energy state and (b) the effect of an acceptor state being ionized.

energy would have to be greater than that of the valence electrons, since the net charge state of the boron atom would now be negative. However, the electron occupying this "empty" position does not have sufficient energy to be in the conduction band, so its energy is far smaller than the conduction-band energy. Figure 4.6b shows how valence electrons may gain a small amount of thermal energy and move about in the crystal. The "empty" position associated with the boron atom becomes occupied, and other valence electron positions become vacated. These other vacated electron positions can be thought of as holes in the semiconductor material.

Figure 4.7 shows the expected energy state of the "empty" position and also the formation of a hole in the valence band. The hole can move through the crystal generating a current, while the negatively charged boron atom is fixed in the crystal. The group III atom accepts an electron from the valence band and so is referred to as an *acceptor impurity atom*. The acceptor atom can generate holes in the valence band without generating electrons in the conduction band. This type of semiconductor material is referred to as a *p-type* material (*p* for the positively charged hole).

The pure single-crystal semiconductor material is called an intrinsic material. Adding controlled amounts of dopant atoms, either donors or acceptors, creates a material called an *extrinsic semiconductor*. An extrinsic semiconductor will have either a preponderance of electrons (n type) or a preponderance of holes (p type).

4.2.2 Ionization Energy

We can calculate the approximate distance of the donor electron from the donor impurity ion, and also the approximate energy required to elevate the donor electron into the conduction band. This energy is referred to as the ionization energy. We will use the Bohr model of the atom for these calculations. The justification for using this model is that the most probable distance of an electron from the nucleus in a hydrogen atom, determined from quantum mechanics, is the same as the Bohr radius. The energy levels in the hydrogen atom determined from quantum mechanics are also the same as obtained from the Bohr theory.

In the case of the donor impurity atom, we may visualize the donor electron orbiting the donor ion, which is embedded in the semiconductor material. We will need to use the permittivity of the semiconductor material in the calculations rather than

the permittivity of free space as is used in the case of the hydrogen atom. We will also use the effective mass of the electron in the calculations.

The analysis begins by setting the coulomb force of attraction between the electron and ion equal to the centripetal force of the orbiting electron. This condition will give a steady orbit. We have

$$\frac{e^2}{4\pi\epsilon r_n^2} = \frac{m^* v^2}{r_n} \tag{4.27}$$

where v is the magnitude of the velocity and r_n is the radius of the orbit. If we assume the angular momentum is also quantized, then we can write

$$m^* r_n v = n\hbar \tag{4.28}$$

where n is a positive integer. Solving for v from Equation (4.28), substituting into Equation (4.27), and solving for the radius, we obtain

$$r_n = \frac{n^2\hbar^2 4\pi\epsilon}{m^* e^2} \tag{4.29}$$

The assumption of the angular momentum being quantized leads to the radius also being quantized.

The Bohr radius is defined as

$$a_0 = \frac{4\pi\epsilon_0\hbar^2}{m_0 e^2} = 0.53\ \text{Å} \tag{4.30}$$

We can normalize the radius of the donor orbital to that of the Bohr radius, which gives

$$\frac{r_n}{a_0} = n^2\epsilon_r\left(\frac{m_0}{m^*}\right) \tag{4.31}$$

where ϵ_r is the relative dielectric constant of the semiconductor material, m_0 is the rest mass of an electron, and m^* is the conductivity effective mass of the electron in the semiconductor.

If we consider the lowest energy state in which $n = 1$, and if we consider silicon in which $\epsilon_r = 11.7$ and the conductivity effective mass is $m^*/m_0 = 0.26$, then we have that

$$\frac{r_1}{a_0} = 45 \tag{4.32}$$

or $r_1 = 23.9$ Å. This radius corresponds to approximately four lattice constants of silicon. Recall that one unit cell in silicon effectively contains eight atoms, so the radius of the orbiting donor electron encompasses many silicon atoms. The donor electron is not tightly bound to the donor atom.

The total energy of the orbiting electron is given by

$$E = T + V \tag{4.33}$$

where T is the kinetic energy and V is the potential energy of the electron. The kinetic energy is

$$T = \frac{1}{2}m^* v^2 \tag{4.34}$$

Using the velocity v from Equation (4.28) and the radius r_n from Equation (4.29), the kinetic energy becomes

$$T = \frac{m^* e^4}{2(n\hbar)^2(4\pi\epsilon)^2} \tag{4.35}$$

The potential energy is

$$V = \frac{-e^2}{4\pi\epsilon r_n} = \frac{-m^* e^4}{(n\hbar)^2(4\pi\epsilon)^2} \tag{4.36}$$

The total energy is the sum of the kinetic and potential energies, so that

$$E = T + V = \frac{-m^* e^4}{2(n\hbar)^2(4\pi\epsilon)^2} \tag{4.37}$$

For the hydrogen atom, $m^* = m_0$ and $\epsilon = \epsilon_0$. The ionization energy of the hydrogen atom in the lowest energy state is then $E = -13.6$ eV. If we consider silicon, the ionization energy is $E = -25.8$ meV, much less than the bandgap energy of silicon. This energy is the approximate ionization energy of the donor atom, or the energy required to elevate the donor electron into the conduction band.

For ordinary donor impurities such as phosphorus or arsenic in silicon or germanium, this hydrogenic model works quite well and gives some indication of the magnitudes of the ionization energies involved. Table 4.3 lists the actual experimentally measured ionization energies for a few impurities in silicon and germanium. Germanium and silicon have different relative dielectric constants and effective masses; thus we expect the ionization energies to differ.

4.2.3 Group III–V Semiconductors

In the previous sections, we have been discussing the donor and acceptor impurities in a group IV semiconductor, such as silicon. The situation in the group III–V

Table 4.3 | Impurity ionization energies in silicon and germanium

	Ionization energy (eV)	
Impurity	**Si**	**Ge**
Donors		
Phosphorus	0.045	0.012
Arsenic	0.05	0.0127
Acceptors		
Boron	0.045	0.0104
Aluminum	0.06	0.0102

Table 4.4 | Impurity ionization energies in gallium arsenide

Impurity	Ionization energy (eV)
Donors	
Selenium	0.0059
Tellurium	0.0058
Silicon	0.0058
Germanium	0.0061
Acceptors	
Beryllium	0.028
Zinc	0.0307
Cadmium	0.0347
Silicon	0.0345
Germanium	0.0404

compound semiconductors, such as gallium arsenide, is more complicated. Group II elements, such as beryllium, zinc, and cadmium, can enter the lattice as substitutional impurities, replacing the group III gallium element to become acceptor impurities. Similarly, group VI elements, such as selenium and tellurium, can enter the lattice substitutionally, replacing the group V arsenic element to become donor impurities. The corresponding ionization energies for these impurities are smaller than for the impurities in silicon. The ionization energies for the donors in gallium arsenide are also smaller than the ionization energies for the acceptors, because of the smaller effective mass of the electron compared to that of the hole.

Group IV elements, such as silicon and germanium, can also be impurity atoms in gallium arsenide. If a silicon atom replaces a gallium atom, the silicon impurity will act as a donor, but if the silicon atom replaces an arsenic atom, then the silicon impurity will act as an acceptor. The same is true for germanium as an impurity atom. Such impurities are called *amphoteric*. Experimentally in gallium arsenide, it is found that germanium is predominantly an acceptor and silicon is predominantly a donor. Table 4.4 lists the ionization energies for the various impurity atoms in gallium arsenide.

TEST YOUR UNDERSTANDING

E4.7 Calculate the radius (normalized to a Bohr radius) of a donor electron in its lowest energy state in GaAs. (Ans. 195.5)

4.3 | THE EXTRINSIC SEMICONDUCTOR

We defined an intrinsic semiconductor as a material with no impurity atoms present in the crystal. An *extrinsic semiconductor* is defined as a semiconductor in which controlled amounts of specific dopant or impurity atoms have been added so that the thermal-equilibrium electron and hole concentrations are different from the intrinsic

carrier concentration. One type of carrier will predominate in an extrinsic semiconductor.

4.3.1 Equilibrium Distribution of Electrons and Holes

Adding donor or acceptor impurity atoms to a semiconductor will change the distribution of electrons and holes in the material. Since the Fermi energy is related to the distribution function, the Fermi energy will change as dopant atoms are added. If the Fermi energy changes from near the midgap value, the density of electrons in the conduction band and the density of holes in the valence band will change. These effects are shown in Figures 4.8 and 4.9. Figure 4.8 shows the case for $E_F > E_{Fi}$ and Figure 4.9 shows the case for $E_F < E_{Fi}$. When $E_F > E_{Fi}$, the electron concentration is larger than the hole concentration, and when $E_F < E_{Fi}$, the hole concentration

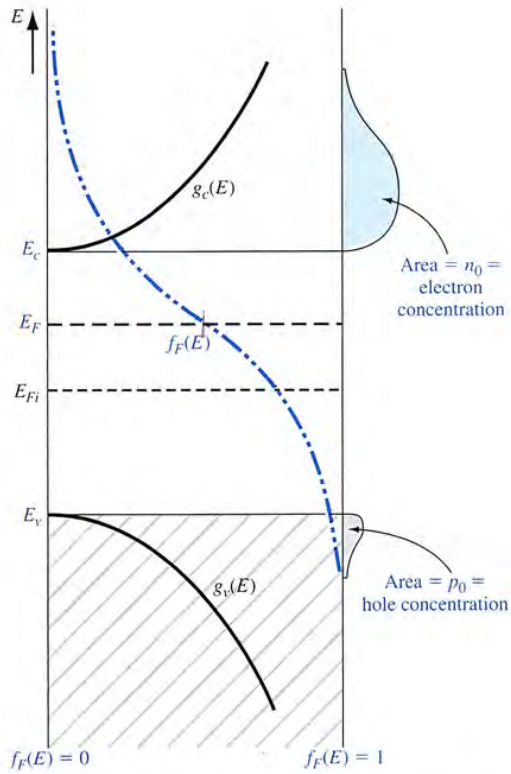

Figure 4.8 | Density of states functions, Fermi–Dirac probability function, and areas representing electron and hole concentrations for the case when E_F is above the intrinsic Fermi energy.

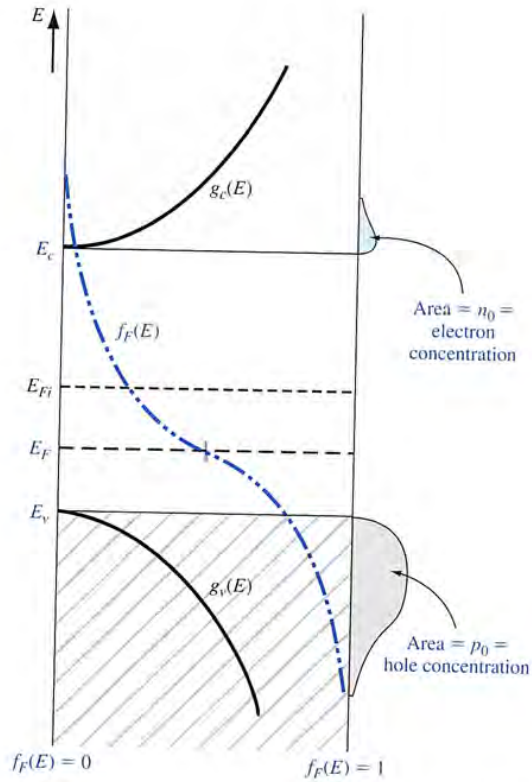

Figure 4.9 | Density of states functions, Fermi–Dirac probability function, and areas representing electron and hole concentrations for the case when E_F is below the intrinsic Fermi energy.

is larger than the electron concentration. When the density of electrons is greater than the density of holes, the semiconductor is n type; donor impurity atoms have been added. When the density of holes is greater than the density of electrons, the semiconductor is p type; acceptor impurity atoms have been added. The Fermi energy level in a semiconductor changes as the electron and hole concentrations change and, again, the Fermi energy changes as donor or acceptor impurities are added. The change in the Fermi level as a function of impurity concentrations will be considered in Section 4.6.

The expressions previously derived for the thermal-equilibrium concentration of electrons and holes, given by Equations (4.11) and (4.19) are general equations for n_0 and p_0 in terms of the Fermi energy. These equations are again given as

$$n_0 = N_c \exp\left[\frac{-(E_c - E_F)}{kT}\right]$$

and

$$p_0 = N_v \exp\left[\frac{-(E_F - E_v)}{kT}\right]$$

As we just discussed, the Fermi energy may vary through the bandgap energy, which will then change the values of n_0 and p_0.

EXAMPLE 4.5

Objective

To calculate the thermal equilibrium concentrations of electrons and holes for a given Fermi energy.

Consider silicon at $T = 300$ K so that $N_c = 2.8 \times 10^{19}$ cm^{-3} and $N_v = 1.04 \times 10^{19}$ cm^{-3}. Assume that the Fermi energy is 0.25 eV below the conduction band. If we assume that the bandgap energy of silicon is 1.12 eV, then the Fermi energy will be 0.87 eV above the valence band.

■ Solution

Using Equation (4.11), we have

$$n_0 = (2.8 \times 10^{19}) \exp\left(\frac{-0.25}{0.0259}\right) = 1.8 \times 10^{15} \text{ cm}^{-3}$$

From Equation (4.19), we can write

$$p_0 = (1.04 \times 10^{19}) \exp\left(\frac{-0.87}{0.0259}\right) = 2.7 \times 10^{4} \text{ cm}^{-3}$$

■ Comment

The change in the Fermi level is actually a function of the donor or acceptor impurity concentrations that are added to the semiconductor. However, this example shows that electron and hole concentrations change by orders of magnitude from the intrinsic carrier concentration as the Fermi energy changes by a few tenths of an electron-volt.

In this example, since $n_0 > p_0$, the semiconductor is n type. In an n-type semiconductor, electrons are referred to as the majority carrier and holes as the minority carrier. By comparing the relative values of n_0 and p_0 in the example, it is easy to see how this designation came about. Similarly, in a p-type semiconductor where $p_0 > n_0$, holes are the majority carrier and electrons are the minority carrier.

We may derive another form of the equations for the thermal-equilibrium concentrations of electrons and holes. If we add and subtract an intrinsic Fermi energy in the exponent of Equation (4.11), we can write

$$n_0 = N_c \exp\left[\frac{-(E_c - E_{Fi}) + (E_F - E_{Fi})}{kT}\right] \tag{4.38a}$$

or

$$n_0 = N_c \exp\left[\frac{-(E_c - E_{Fi})}{kT}\right] \exp\left[\frac{(E_F - E_{Fi})}{kT}\right] \tag{4.38b}$$

The intrinsic carrier concentration is given by Equation (4.20) as

$$n_i = N_c \exp\left[\frac{-(E_c - E_{Fi})}{kT}\right]$$

so that the thermal-equilibrium electron concentration can be written as

$$\boxed{n_0 = n_i \exp\left[\frac{E_F - E_{Fi}}{kT}\right]} \tag{4.39}$$

Similarly, if we add and subtract an intrinsic Fermi energy in the exponent of Equation (4.19), we will obtain

$$\boxed{p_0 = n_i \exp\left[\frac{-(E_F - E_{Fi})}{kT}\right]} \tag{4.40}$$

As we will see, the Fermi level changes when donors and acceptors are added, but Equations (4.39) and (4.40) show that, as the Fermi level changes from the intrinsic Fermi level, n_0 and p_0 change from the n_i value. If $E_F > E_{Fi}$, then we will have $n_0 > n_i$ and $p_0 < n_i$. One characteristic of an n-type semiconductor is that $E_F > E_{Fi}$ so that $n_0 > p_0$. Similarly, in a p-type semiconductor, $E_F < E_{Fi}$ so that $p_0 > n_i$ and $n_0 < n_i$; thus $p_0 > n_0$.

We can see the functional dependence of n_0 and p_0 with E_F in Figures 4.8 and 4.9. As E_F moves above or below E_{Fi}, the overlapping probability function with the density of states functions in the conduction band and valence band changes. As E_F moves above E_{Fi}, the probability function in the conduction band increases, while the probability, $1 - f_F(E)$, of an empty state (hole) in the valence band decreases. As E_F moves below E_{Fi}, the opposite occurs.

4.3.2 The $n_0 p_0$ Product

We may take the product of the general expressions for n_0 and p_0 as given in Equations (4.11) and (4.19), respectively. The result is

$$n_0 p_0 = N_c N_v \exp\left[\frac{-(E_c - E_F)}{kT}\right] \exp\left[\frac{-(E_F - E_v)}{kT}\right] \tag{4.41}$$

which may be written as

$$n_0 p_0 = N_c N_v \exp\left[\frac{-E_g}{kT}\right] \tag{4.42}$$

As Equation (4.42) was derived for a general value of Fermi energy, the values of n_0 and p_0 are not necessarily equal. However, Equation (4.42) is exactly the same as Equation (4.23), which we derived for the case of an intrinsic semiconductor. We

which yields

$$N_d = 1.39 \times 10^{15} \text{ cm}^{-3}$$

■ Comment

If the temperature remains less than $T = 550$ K, then the intrinsic carrier concentration will contribute less than 5 percent of the total electron concentration for this donor impurity concentration.

Equations (4.60) and (4.62) are used to calculate the majority carrier electron concentration in an n-type semiconductor and majority carrier hole concentration in a p-type semiconductor, respectively. The minority carrier hole concentration in an n-type semiconductor could, theoretically, be calculated from Equation (4.62). However, we would be subtracting two numbers on the order of 10^{16} cm^{-3}, for example, to obtain a number on the order of 10^4 cm^{-3}, which from a practical point of view is not possible. The minority carrier concentrations are calculated from $n_0 p_0 = n_i^2$ once the majority carrier concentration has been determined.

TEST YOUR UNDERSTANDING

E4.11 Consider a compensated GaAs semiconductor at $T = 300$ K doped at $N_d = 5 \times 10^{15}$ cm^{-3} and $N_a = 2 \times 10^{16}$ cm^{-3}. Calculate the thermal equilibrium electron and hole concentrations. (Ans. $p_0 = 1.5 \times 10^{16}$ cm^{-3}, $n_0 = 2.16 \times 10^{-4}$ cm^{-3})

E4.12 Silicon is doped at $N_d = 10^{15}$ cm^{-3} and $N_a = 0$. (*a*) Plot the concentration of electrons versus temperature over the range $300 \le T \le 600$ K. (*b*) Calculate the temperature at which the electron concentration is equal to 1.1×10^{15} cm^{-3}. (Ans. $T \approx 552$ K)

4.6 | POSITION OF FERMI ENERGY LEVEL

We discussed qualitatively in Section 4.3.1 how the electron and hole concentrations change as the Fermi energy level moves through the bandgap energy. Then, in Section 4.5, we calculated the electron and hole concentrations as a function of donor and acceptor impurity concentrations. We can now determine the position of the Fermi energy level as a function of the doping concentrations and as a function of temperature. The relevance of the Fermi energy level will be further discussed after the mathematical derivations.

4.6.1 Mathematical Derivation

The position of the Fermi energy level within the bandgap can be determined by using the equations already developed for the thermal-equilibrium electron and hole concentrations. If we assume the Boltzmann approximation to be valid, then from Equation (4.11) we have $n_0 = N_c \exp[-(E_c - E_F)/kT]$. We can solve for $E_c - E_F$

from this equation and obtain

$$\boxed{E_c - E_F = kT \ln\left(\frac{N_c}{n_0}\right)} \tag{4.63}$$

where n_0 is given by Equation (4.60). If we consider an n-type semiconductor in which $N_d \gg n_i$, then $n_0 \approx N_d$, so that

$$E_c - E_F = kT \ln\left(\frac{N_c}{N_d}\right) \tag{4.64}$$

The distance between the bottom of the conduction band and the Fermi energy is a logarithmic function of the donor concentration. As the donor concentration increases, the Fermi level moves closer to the conduction band. Conversely, if the Fermi level moves closer to the conduction band, then the electron concentration in the conduction band is increasing. We may note that if we have a compensated semiconductor, then the N_d term in Equation (4.64) is simply replaced by $N_d - N_a$, or the net effective donor concentration.

DESIGN EXAMPLE 4.13

Objective

To determine the required donor impurity concentration to obtain a specified Fermi energy.

Silicon at $T = 300$ K contains an acceptor impurity concentration of $N_a = 10^{16}$ cm^{-3}. Determine the concentration of donor impurity atoms that must be added so that the silicon is n type and the Fermi energy is 0.20 eV below the conduction band edge.

■ Solution

From Equation (4.64), we have

$$E_c - E_F = kT \ln\left(\frac{N_c}{N_d - N_a}\right)$$

which can be rewritten as

$$N_d - N_a = N_c \exp\left[\frac{-(E_c - E_F)}{kT}\right]$$

Then

$$N_d - N_a = 2.8 \times 10^{19} \exp\left[\frac{-0.20}{0.0259}\right] = 1.24 \times 10^{16} \text{ cm}^{-3}$$

or

$$N_d = 1.24 \times 10^{16} + N_a = 2.24 \times 10^{16} \text{ cm}^{-3}$$

■ Comment

A compensated semiconductor can be fabricated to provide a specific Fermi energy level.

We may develop a slightly different expression for the position of the Fermi level. We had from Equation (4.39) that $n_0 = n_i \exp[(E_F - E_{Fi})/kT]$. We can solve for $E_F - E_{Fi}$ as

$$\boxed{E_F - E_{Fi} = kT \ln\left(\frac{n_0}{n_i}\right)} \tag{4.65}$$

Equation (4.65) can be used specifically for an n-type semiconductor, where n_0 is given by Equation (4.60), to find the difference between the Fermi level and the intrinsic Fermi level as a function of the donor concentration. We may note that, if the net effective donor concentration is zero, that is, $N_d - N_a = 0$, then $n_0 = n_i$ and $E_F = E_{Fi}$. A completely compensated semiconductor has the characteristics of an intrinsic material in terms of carrier concentration and Fermi level position.

We can derive the same types of equations for a p-type semiconductor. From Equation (4.19), we have $p_0 = N_v \exp[-(E_F - E_v)/kT]$, so that

$$\boxed{E_F - E_v = kT \ln\left(\frac{N_v}{p_0}\right)} \tag{4.66}$$

If we assume that $N_a \gg n_i$, then Equation (4.66) can be written as

$$E_F - E_v = kT \ln\left(\frac{N_v}{N_a}\right) \tag{4.67}$$

The distance between the Fermi level and the top of the valence-band energy for a p-type semiconductor is a logarithmic function of the acceptor concentration: as the acceptor concentration increases, the Fermi level moves closer to the valence band. Equation (4.67) still assumes that the Boltzmann approximation is valid. Again, if we have a compensated p-type semiconductor, then the N_a term in Equation (4.67) is replaced by $N_a - N_d$, or the net effective acceptor concentration.

We can also derive an expression for the relationship between the Fermi level and the intrinsic Fermi level in terms of the hole concentration. We have from Equation (4.40) that $p_0 = n_i \exp[-(E_F - E_{Fi})/kT]$, which yields

$$\boxed{E_{Fi} - E_F - kT \ln\left(\frac{p_0}{n_i}\right)} \tag{4.68}$$

Equation (4.68) can be used to find the difference between the intrinsic Fermi level and the Fermi energy in terms of the acceptor concentration. The hole concentration p_0 in Equation (4.68) is given by Equation (4.62).

We may again note from Equation (4.65) that, for an n-type semiconductor, $n_0 > n_i$ and $E_F > E_{Fi}$. The Fermi level for an n-type semiconductor is above E_{Fi}. For a p-type semiconductor, $p_0 > n_i$, and from Equation (4.68) we see that

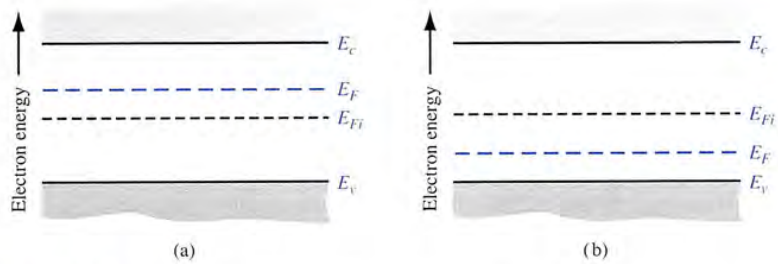

Figure 4.17 | Position of Fermi level for an (a) n-type ($N_d > N_a$) and (b) p-type ($N_d > N_a$) semiconductor.

$E_{Fi} > E_F$. The Fermi level for a p-type semiconductor is below E_{Fi}. These results are shown in Figure 4.17.

4.6.2 Variation of E_F with Doping Concentration and Temperature

We may plot the position of the Fermi energy level as a function of the doping concentration. Figure 4.18 shows the Fermi energy level as a function of donor concentration (n type) and as a function of acceptor concentration (p type) for silicon at $T = 300$ K. As the doping levels increase, the Fermi energy level moves closer to the conduction band for the n-type material and closer to the valence band for the p-type material. Keep in mind that the equations for the Fermi energy level that we have derived assume that the Boltzmann approximation is valid.

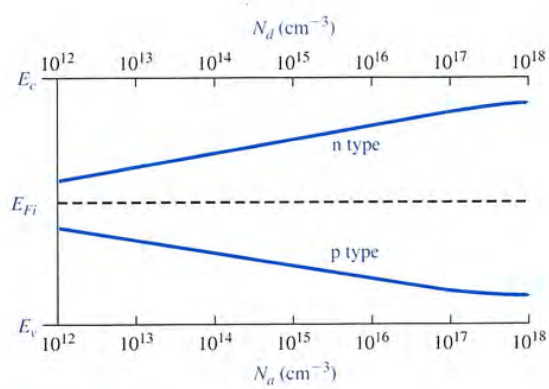

Figure 4.18 | Position of Fermi level as a function of donor concentration (n type) and acceptor concentration (p type).

Objective | EXAMPLE 4.14

To determine the Fermi-level position and the maximum doping at which the Boltzmann approximation is still valid.

Consider p-type silicon, at $T = 300$ K, doped with boron. We may assume that the limit of the Boltzmann approximation occurs when $E_F - E_a = 3kT$. (See Section 4.1.2.)

■ Solution

From Table 4.3, we find the ionization energy is $E_a - E_v = 0.045$ eV for boron in silicon. If we assume that $E_{Fi} \approx E_{\text{midgap}}$, then from Equation (4.68), the position of the Fermi level at the maximum doping is given by

$$E_{Fi} - E_F = \frac{E_g}{2} - (E_a - E_v) - (E_F - E_a) = kT \ln\left(\frac{N_a}{n_i}\right)$$

or

$$0.56 - 0.045 - 3(0.0259) = 0.437 = (0.0259) \ln\left(\frac{N_a}{n_i}\right)$$

We can then solve for the doping as

$$N_a = n_i \exp\left(\frac{0.437}{0.0259}\right) = 3.2 \times 10^{17} \text{ cm}^{-3}$$

■ Comment

If the acceptor (or donor) concentration in silicon is greater than approximately 3×10^{17} cm^{-3}, then the Boltzmann approximation of the distribution function becomes less valid and the equations for the Fermi-level position are no longer quite as accurate.

TEST YOUR UNDERSTANDING

E4.13 Determine the position of the Fermi level with respect to the valence band energy in p type GaAs at $T = 300$ K. The doping concentrations are $N_a = 5 \times 10^{16}$ cm^{-3} and $N_d = 4 \times 10^{15}$ cm^{-3}. (Ans. $E_F - E_a = 0.130$ eV)

E4.14 Calculate the position of the Fermi energy level in n-type silicon at $T = 300$ K with respect to the intrinsic Fermi energy level. The doping concentrations are $N_d = 2 \times 10^{17}$ cm^{-3} and $N_a = 3 \times 10^{16}$ cm^{-3}. (Ans. $E_F - E_{Fi} = 0.421$ eV)

The intrinsic carrier concentration n_i, in Equations (4.65) and (4.68), is a strong function of temperature, so that E_F is a function of temperature also. Figure 4.19 shows the variation of the Fermi energy level in silicon with temperature for several donor and acceptor concentrations. As the temperature increases, n_i increases, and E_F moves closer to the intrinsic Fermi level. At high temperature, the semiconductor material begins to lose its extrinsic characteristics and begins to behave more like an intrinsic semiconductor. At the very low temperature, freeze-out occurs; the Boltzmann approximation is no longer valid and the equations we derived for the

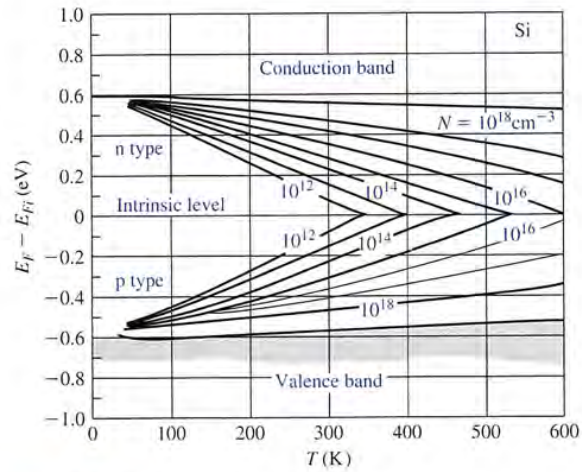

Figure 4.19 | Position of Fermi level as a function of temperature for various doping concentrations. *(From Sze [13].)*

Fermi-level position no longer apply. At the low temperature where freeze-out occurs, the Fermi level goes above E_d for the n-type material and below E_a for the p-type material. At absolute zero degrees, all energy states below E_F are full and all energy states above E_F are empty.

4.6.3 Relevance of the Fermi Energy

We have been calculating the position of the Fermi energy level as a function of doping concentrations and temperature. This analysis may seem somewhat arbitrary and fictitious. However, these relations do become significant later in our discussion of pn junctions and the other semiconductor devices we consider. An important point is that, in thermal equilibrium, the Fermi energy level is a constant throughout a system. We will not prove this statement, but we can intuitively see its validity by considering the following example.

Suppose we have a particular material, A, whose electrons are distributed in the energy states of an allowed band as shown in Figure 4.20a. Most of the energy states below E_{FA} contain electrons and most of the energy states above E_{FA} are empty of electrons. Consider another material, B, whose electrons are distributed in the energy states of an allowed band as shown in Figure 4.20b. The energy states below E_{FB} are mostly full and the energy states above E_{FB} are mostly empty. If these two materials are brought into intimate contact, the electrons in the entire system will tend to seek the lowest possible energy. Electrons from material A will flow into the lower energy states of material B, as indicated in Figure 4.20c, until thermal equilibrium is reached. Thermal equilibrium occurs when the distribution of electrons, as

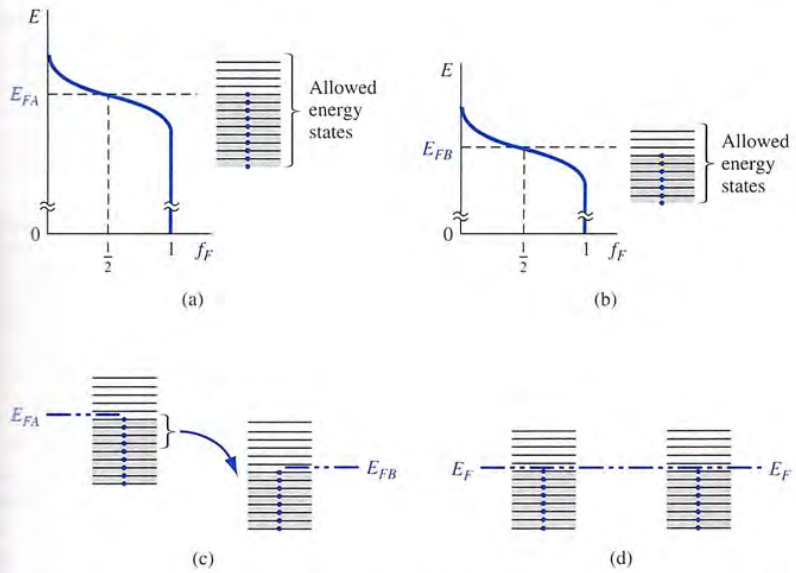

Figure 4.20 | The Fermi energy of (a) material A in thermal equilibrium, (b) material B in thermal equilibrium, (c) materials A and B at the instant they are placed in contact, and (d) materials A and B in contact at thermal equilibrium.

a function of energy, is the same in the two materials. This equilibrium state occurs when the Fermi energy is the same in the two materials as shown in Figure 4.20d. The Fermi energy, important in the physics of the semiconductor, also provides a good pictorial representation of the characteristics of the semiconductor materials and devices.

4.7 | SUMMARY

- The concentration of electrons in the conduction band is the integral over the conduction band energy of the product of the density of states function in the conduction band and the Fermi–Dirac probability function.
- The concentration of holes in the valence band is the integral over the valence band energy of the product of the density of states function in the valence band and the probability of a state being empty, which is $[1 - f_F(E)]$.
- Using the Maxwell–Boltzmann approximation, the thermal equilibrium concentration of electrons in the conduction band is given by

$$n_0 = N_c \exp\left[\frac{-(E_c - E_F)}{kT}\right]$$

where N_c is the effective density of states in the conduction band.

CHAPTER 11

Fundamentals of the Metal–Oxide–Semiconductor Field-Effect Transistor

PREVIEW

The fundamental physics of the **M**etal–**O**xide–**S**emiconductor **F**ield-**E**ffect **T**ransistor (MOSFET) is developed in this chapter. Although the bipolar transistor was discussed in the last chapter, the material in this chapter presumes a knowledge only of the semiconductor material properties and characteristics of the pn junction.

The MOSFET, in conjunction with other circuit elements, is capable of voltage gain and signal-power gain. The MOSFET is also used extensively in digital circuit applications where, because of its relatively small size, thousands of devices can be fabricated in a single integrated circuit. The MOSFET is, without doubt, the core of integrated circuit design at the present time.

The MOS designation is implicity used only for the metal–silicon dioxide (SiO_2)–silicon system. The more general terminology is metal–insulator–semiconductor (MIS), where the insulator is not necessarily silicon dioxide and the semiconductor is not necessarily silicon. We will use the MOS system throughout this chapter although the same basic physics applies to the MIS system.

The heart of the MOSFET is a metal–oxide–semiconductor structure known as an MOS capacitor. The energy bands in the semiconductor near the oxide–semiconductor interface bend as a voltage is applied across the MOS capacitor. The position of the conduction and valence bands relative to the Fermi level at the oxide–semiconductor interface is a function of the MOS capacitor voltage, so that the characteristics of the semiconductor surface can be inverted from p-type to n-type, or from n-type to p-type, by applying the proper voltage. The operation and characteristics of the MOSFET are dependent on this inversion and the creation of

an inversion charge density at the semiconductor surface. The threshold voltage is defined as the applied gate voltage required to create the inversion layer charge and is one of the important parameters of the MOSFET.

The various types of MOSFETs are examined and a qualitative discussion of the current–voltage characteristics is initially presented. A mathematical derivation of the current–voltage relation is then covered in detail. The frequency response and limitations of the MOSFET are also considered.

Although we have not discussed fabrication processes in any detail in this text, there is an MOS technology that should be considered, since it directly influences the characteristics and properties of the MOS devices and circuits. We will consider the complementary MOS (CMOS) process. The discussion of this technology will be brief, but should provide a good base for further in-depth study. ■

11.1 | THE TWO-TERMINAL MOS STRUCTURE

The heart of the MOSFET is the metal–oxide–semiconductor capacitor shown in Figure 11.1. The metal may be aluminum or some other type of metal, although in many cases, it is actually a high-conductivity polycrystalline silicon that has been deposited on the oxide; however, the term metal is usually still used. The parameter t_{ox} in the figure is the thickness of the oxide and ϵ_{ox} is the permittivity of the oxide.

11.1.1 Energy-Band Diagrams

The physics of the MOS structure can be more easily explained with the aid of the simple parallel-plate capacitor. Figure 11.2a shows a parallel-plate capacitor with the top plate at a negative voltage with respect to the bottom plate. An insulator material separates the two plates. With this bias, a negative charge exists on the top plate, a positive charge exists on the bottom plate, and an electric field is induced between the two plates as shown. The capacitance per unit area for this geometry is

$$C' = \frac{\epsilon}{d} \tag{11.1}$$

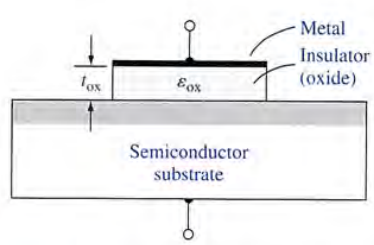

Figure 11.1 | The basic MOS capacitor structure.

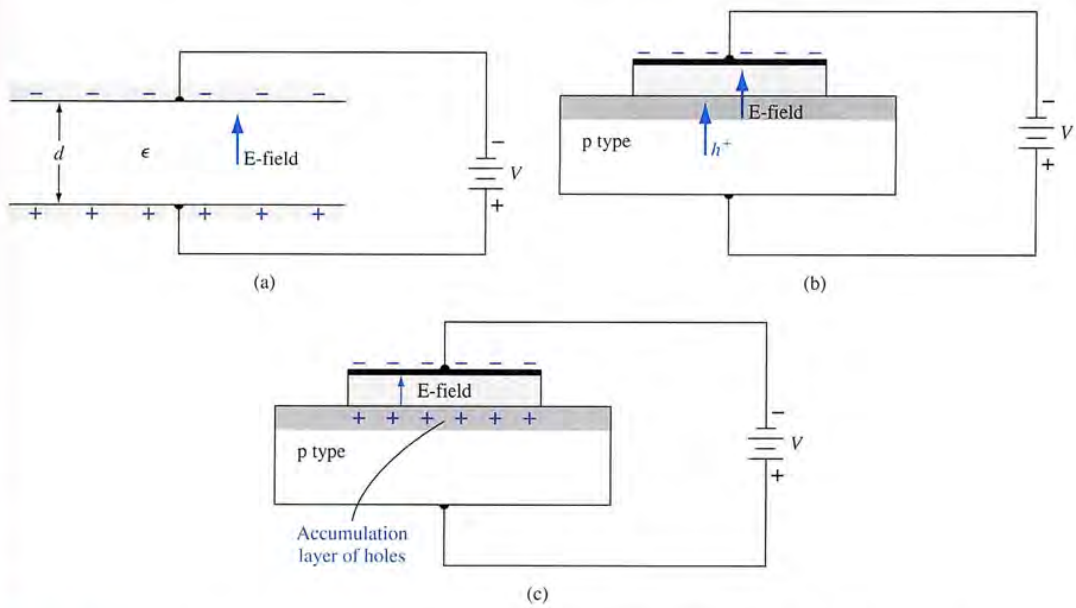

Figure 11.2 | (a) A parallel-plate capacitor showing the electric field and conductor charges. (b) A corresponding MOS capacitor with a negative gate bias showing the electric field and charge flow. (c) The MOS capacitor with an accumulation layer of holes.

where ϵ is the permittivity of the insulator and d is the distance between the two plates. The magnitude of the charge per unit area on either plate is

$$Q' = C'V \tag{11.2}$$

where the prime indicates charge or capacitance per unit area. The magnitude of the electric field is

$$\mathrm{E} = \frac{V}{d} \tag{11.3}$$

Figure 11.2b shows an MOS capacitor with a p-type semiconductor substrate. The top metal gate is at a negative voltage with respect to the semiconductor substrate. From the example of the parallel-plate capacitor, we can see that a negative charge will exist on the top metal plate and an electric field will be induced with the direction shown in the figure. If the electric field were to penetrate into the semiconductor, the majority carrier holes would experience a force toward the oxide–semiconductor interface. Figure 11.2c shows the equilibrium distribution of charge in the MOS capacitor with this particular applied voltage. An *accumulation layer* of holes in the oxide–semiconductor junction corresponds to the positive charge on the bottom "plate" of the MOS capacitor.

Figure 11.3a shows the same MOS capacitor in which the polarity of the applied voltage is reversed. A positive charge now exists on the top metal plate and the induced electric field is in the opposite direction as shown. If the electric field penetrates the semiconductor in this case, majority carrier holes will experience a force away from the oxide–semiconductor interface. As the holes are pushed away from the interface, a negative space charge region is created because of the fixed ionized acceptor atoms. The negative charge in the induced depletion region corresponds to the negative charge on the bottom "plate" of the MOS capacitor. Figure 11.3b shows the equilibrium distribution of charge in the MOS capacitor with this applied voltage.

The energy-band diagram of the MOS capacitor with the p-type substrate, for the case when a negative voltage is applied to the top metal gate, is shown in Figure 11.4a. The valence-band edge is closer to the Fermi level at the oxide–semiconductor interface than in the bulk material, which implies that there is an accumulation of holes. The semiconductor surface appears to be more p-type than the bulk material. The Fermi level is a constant in the semiconductor since the MOS system is in thermal equilibrium and there is no current through the oxide.

Figure 11.4b shows the energy-band diagram of the MOS system when a positive voltage is applied to the gate. The conduction and valence band edges bend as

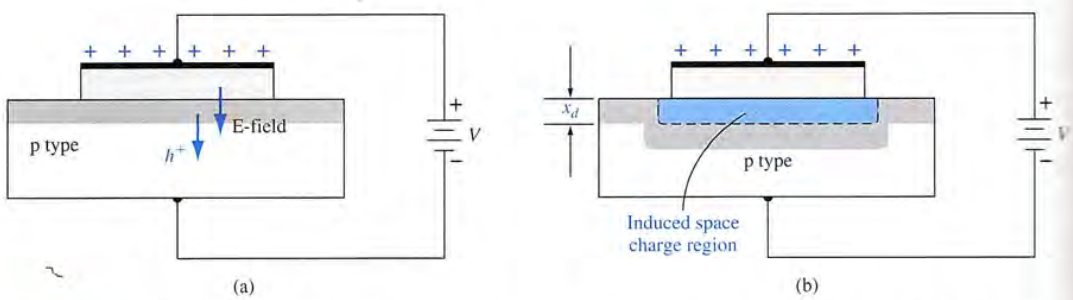

Figure 11.3 | The MOS capacitor with a moderate positive gate bias, showing (a) the electric field and charge flow and (b) the induced space charge region.

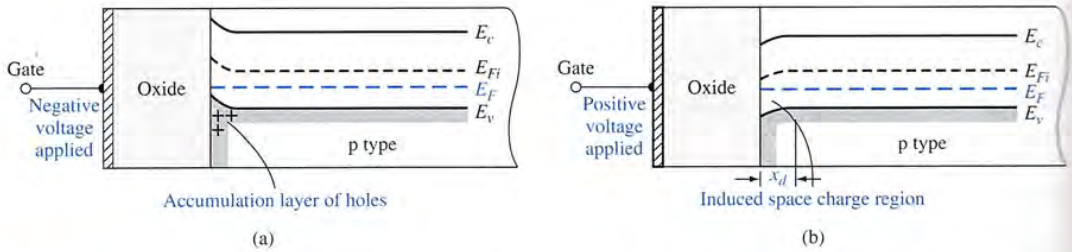

Figure 11.4 | The energy-band diagram of an MOS capacitor with a p-type substrate for (a) a negative gate bias and (b) a moderate positive gate bias.

shown in the figure, indicating a space charge region similar to that in a pn junction. The conduction band and intrinsic Fermi levels move closer to the Fermi level. The induced space charge width is x_d.

Now consider the case when a still larger positive voltage is applied to the top metal gate of the MOS capacitor. We expect the induced electric field to increase in magnitude and the corresponding positive and negative charges on the MOS capacitor to increase. A larger negative charge in the MOS capacitor implies a larger induced space charge region and more band bending. Figure 11.5 shows such a condition. The intrinsic Fermi level at the surface is now below the Fermi level; thus, the conduction band is closer to the Fermi level than the valence band is. This result implies that the surface in the semiconductor adjacent to the oxide-semiconductor interface is n type. By applying a sufficiently large positive gate voltage, we have inverted the surface of the semiconductor from a p-type to an n-type semiconductor. We have created an *inversion layer* of electrons at the oxide-semiconductor interface.

In the MOS capacitor structure that we have just considered, we assumed a p-type semiconductor substrate. The same type of energy-band diagrams can be constructed for an MOS capacitor with an n-type semiconductor substrate. Figure 11.6a shows the MOS capacitor structure with a positive voltage applied to the top gate terminal. A positive charge exists on the top gate and an electric field is induced with the direction shown in the figure. An accumulation layer of electrons will be induced in

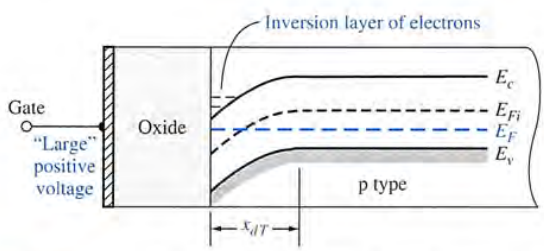

Figure 11.5 | The energy-band diagram of the MOS capacitor with a p-type substrate for a "large" positive gate bias.

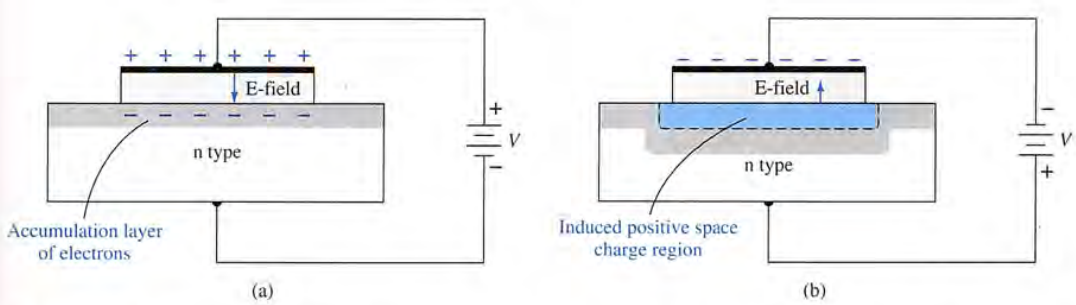

Figure 11.6 | The MOS capacitor with an n-type substrate for (a) a positive gate bias and (b) a moderate negative gate bias.

the n-type substrate. The case when a negative voltage is applied to the top gate is shown in Figure 11.6b. A positive space charge region is induced in the n-type semiconductor in this situation.

The energy-band diagrams for this MOS capacitor with the n-type substrate are shown in Figure 11.7. Figure 11.7a shows the case when a positive voltage is applied to the gate and an accumulation layer of electrons is formed. Figure 11.7b shows the positive space charge region induced by an applied negative gate voltage; in it the conduction and valence band energies bend upward. Figure 11.7c shows the energy bands when a larger negative voltage is applied to the gate. The conduction and valence bands are bent even more and the intrinsic Fermi level has moved above the Fermi level so that the valence band is closer to the Fermi level than the

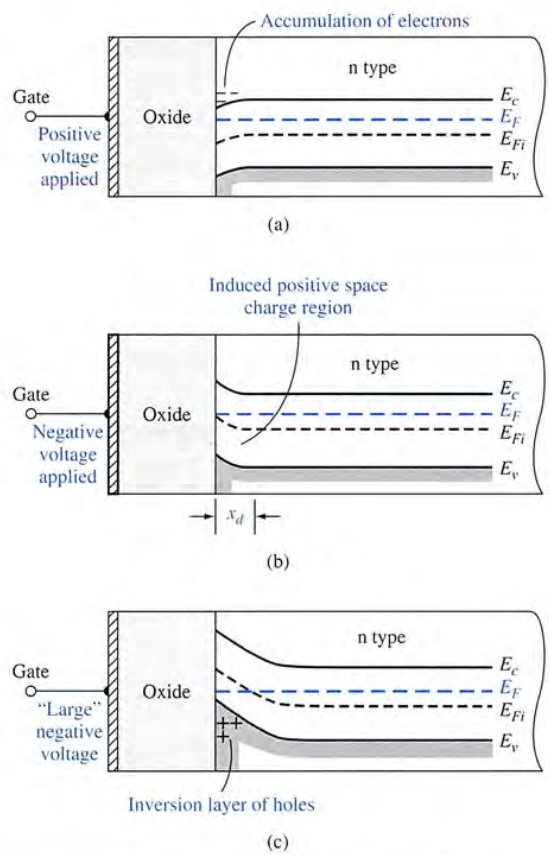

Figure 11.7 | The energy-band diagram of the MOS capacitor with an n-type substrate for (a) a positive gate bias, (b) a moderate negative bias, and (c) a "large" negative gate bias.

conduction band is. This result implies that the semiconductor surface adjacent to the oxide-semiconductor interface is p type. By applying a sufficiently large negative voltage to the gate of the MOS capacitor, the semiconductor surface has been inverted from n type to p type. An inversion layer of holes has been induced at the oxide-semiconductor interface.

11.1.2 Depletion Layer Thickness

We may calculate the width of the induced space charge region adjacent to the oxide-semiconductor interface. Figure 11.8 shows the space charge region in a p-type semiconductor substrate. The potential ϕ_{fp} is the difference (in volts) between E_{Fi} and E_F and is given by

$$\boxed{\phi_{fp} = V_t \ln\left(\frac{N_a}{n_i}\right)} \tag{11.4}$$

where N_a is the acceptor doping concentration and n_i is the intrinsic carrier concentration.

The potential ϕ_s is called the surface potential; it is the difference (in volts) between E_{Fi} measured in the bulk semiconductor and E_{Fi} measured at the surface. The surface potential is the potential difference across the space charge layer. The space charge width can now be written in a form similar to that of a one-sided pn junction. We can write that

$$x_d = \left(\frac{2\epsilon_s \phi_s}{eN_a}\right)^{1/2} \tag{11.5}$$

where ϵ_s is the permittivity of the semiconductor. Equation (11.5) assumes that the abrupt depletion approximation is valid.

Figure 11.9 shows the energy bands for the case in which $\phi_s = 2\phi_{fp}$. The Fermi level at the surface is as far above the intrinsic level as the Fermi level is below the

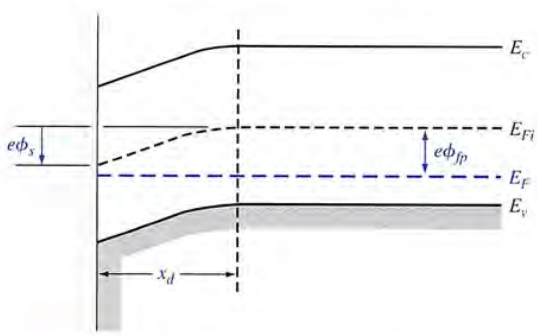

Figure 11.8 | The energy-band diagram in the p-type semiconductor, indicating surface potential.

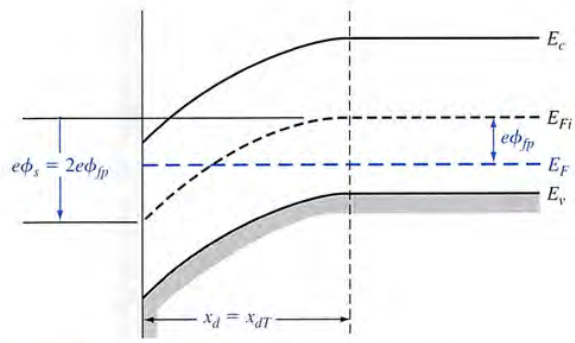

Figure 11.9 | The energy-band diagram in the p-type semiconductor at the threshold inversion point.

intrinsic level in the bulk semiconductor. The electron concentration at the surface is the same as the hole concentration in the bulk material. This condition is known as the *threshold inversion point.* The applied gate voltage creating this condition is known as the *threshold voltage.* If the gate voltage increases above this threshold value, the conduction band will bend slightly closer to the Fermi level, but the change in the conduction band at the surface is now only a slight function of gate voltage. The electron concentration at the surface, however, is an exponential function of the surface potential. The surface potential may increase by a few (kT/e) volts, which will change the electron concentration by orders of magnitude, but the space charge width changes only slightly. In this case, then, the space charge region has essentially reached a maximum width.

The maximum space charge width, x_{dT}, at this inversion transition point can be calculated from Equation (11.5) by setting $\phi_s = 2\phi_{fp}$. Then

$$\boxed{x_{dT} = \left(\frac{4\epsilon_s \phi_{fp}}{eN_a}\right)^{1/2}} \tag{11.6}$$

EXAMPLE 11.1 | Objective

To calculate the maximum space charge width given a particular semiconductor doping concentration.

Consider silicon at $T = 300$ K doped to $N_a = 10^{16}$ cm^{-3}. The intrinsic carrier concentration is $n_i = 1.5 \times 10^{10}$ cm^{-3}.

■ **Solution**

From Equation (11.4), we have

$$\phi_{fp} = V_t \ln\left(\frac{N_a}{n_i}\right) = (0.0259)\ln\left(\frac{10^{16}}{1.5 \times 10^{10}}\right) = 0.347 \text{ V}$$

Then the maximum space charge width is

$$x_{dT} = \left[\frac{4\epsilon_s \phi_{fp}}{eN_a}\right]^{1/2} = \left[\frac{4(11.7)(8.85 \times 10^{-14})(0.347)}{(1.6 \times 10^{-19})(10^{16})}\right]^{1/2}$$

or

$$x_{dT} = 0.30 \times 10^{-4} \text{ cm} = 0.30\ \mu\text{m}$$

■ Comment

The maximum induced space charge width is on the same order of magnitude as pn junction space charge widths.

We have been considering a p-type semiconductor substrate. The same maximum induced space charge region width occurs in an n-type substrate. Figure 11.10 is the energy-band diagram at the threshold voltage with an n-type substrate. We can write

$$\boxed{\phi_{fn} = V_t \ln\left(\frac{N_d}{n_i}\right)} \tag{11.7}$$

and

$$\boxed{x_{dT} = \left(\frac{4\epsilon_s \phi_{fn}}{eN_d}\right)^{1/2}} \tag{11.8}$$

Note that we are always assuming the parameters ϕ_{fp} and ϕ_{fn} to be positive quantities.

Figure 11.11 is a plot of x_{dT} at $T = 300$ K as a function of doping concentration in silicon. The semiconductor doping can be either n-type or p-type.

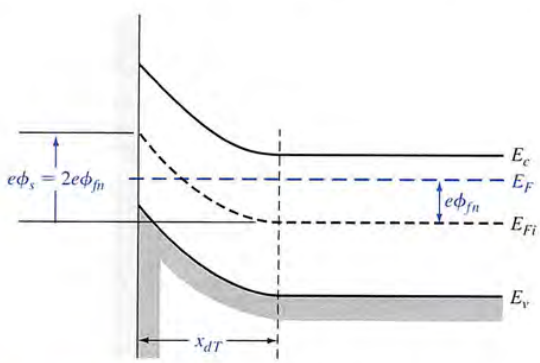

Figure 11.10 | The energy-band diagram in the n-type semiconductor at the threshold inversion point.

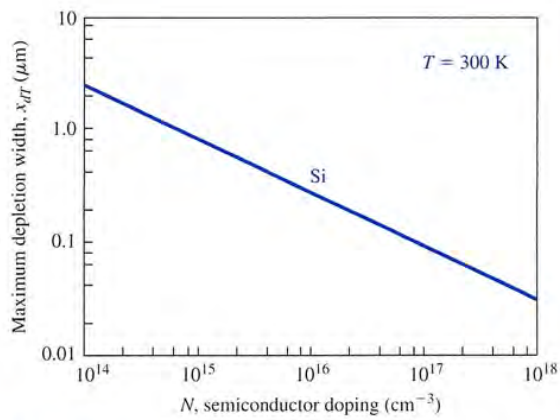

Figure 11.11 | Maximum induced space charge region width versus semiconductor doping.

TEST YOUR UNDERSTANDING

E11.1 (*a*) Consider as oxide-to-p-type silicon junction at $T = 300$ K. The impurity doping concentration in the silicon is $N_a = 3 \times 10^{16}$ cm^{-3}. Calculate the maximum space-charge width in the silicon. (*b*) Repeat part (*a*) for an impurity concentration of $N_a = 10^{15}$ cm^{-3}. [(*a*) 0.180 μm; (*b*) 0.863 μm]

E11.2 Consider an oxide-to-n-type silicon junction at $T = 300$ K. The impurity doping concentration in the silicon is $N_d = 8 \times 10^{15}$ cm^{-3}. Calculate the maximum space-charge width in the silicon. (0.33 μm)

11.1.3 Work Function Differences

We have been concerned, so far, with the energy-band diagrams of the semiconductor material. Figure 11.12a shows the energy levels in the metal, silicon dioxide, and silicon relative to the vacuum level. The metal work function is ϕ_m and the electron affinity is χ. The parameter χ_i is the oxide electron affinity and, for silicon dioxide, $\chi_i = 0.9$ V.

Figure 11.12b shows the energy-band diagram of the entire metal–oxide–semiconductor structure with zero gate voltage applied. The Fermi level is a constant through the entire system at thermal equilibrium. We may define ϕ'_m as a modified metal work function—the potential required to inject an electron from the metal into the conduction band of the oxide. Similarly, χ' is defined as a modified electron affinity. The voltage V_{ox0} is the potential drop across the oxide for zero applied gate voltage and is not necessarily zero because of the difference between ϕ_m and χ. The potential ϕ_{s0} is the surface potential for this case.

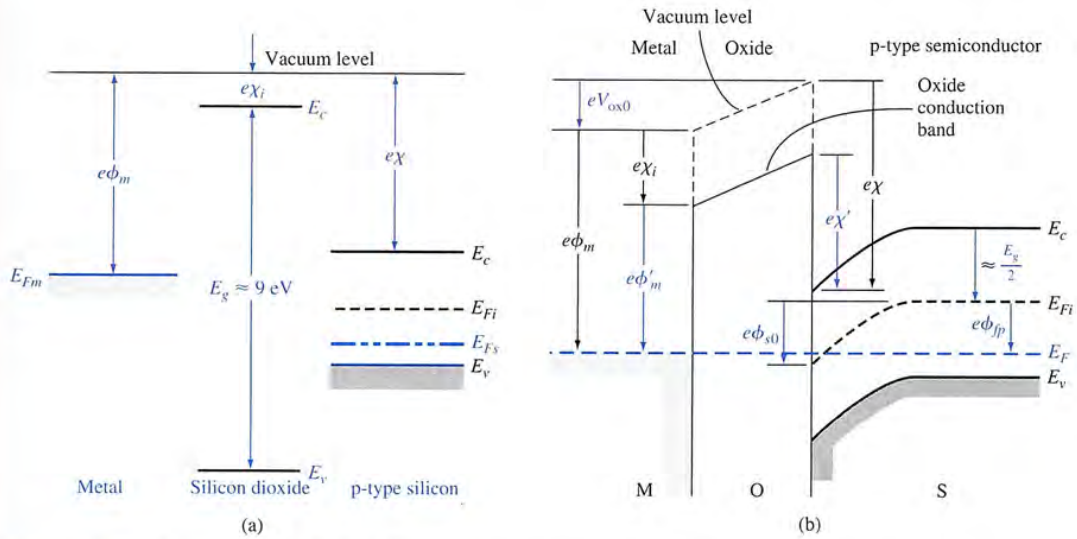

Figure 11.12 | (a) Energy levels in an MOS system prior to contact and (b) energy-band diagram through the MOS structure in thermal equilibrium after contact.

If we sum the energies from the Fermi level on the metal side to the Fermi level on the semiconductor side, we have

$$e\phi'_m + eV_{ox0} = e\chi' + \frac{E_g}{2} - e\phi_{s0} + e\phi_{fp} \tag{11.9}$$

Equation (11.9) can be rewritten as

$$V_{ox0} + \phi_{s0} = -\left[\phi'_m - \left(\chi' + \frac{E_g}{2e} + \phi_{fp}\right)\right] \tag{11.10}$$

We can define a potential ϕ_{ms} as

$$\boxed{\phi_{ms} \equiv \left[\phi'_m - \left(\chi' + \frac{E_g}{2e} + \phi_{fp}\right)\right]} \tag{11.11}$$

which is known as the metal–semiconductor work function difference.

Objective | **EXAMPLE 11.2**

To calculate the metal–semiconductor work function difference ϕ_{ms} for a given MOS system and semiconductor doping.

For an aluminium–silicon dioxide junction, $\phi'_m = 3.20$ V and for a silicon–silicon dioxide junction, $\chi' = 3.25$ V. We may assume that $E_g = 1.11$ eV. Let the p-type doping be $N_a = 10^{14}$ cm^{-3}.

■ Solution

For silicon at $T = 300$ K, we may calculate ϕ_{fp} as

$$\phi_{fp} = V_t \ln\left(\frac{N_a}{n_i}\right) = (0.0259)\ln\left(\frac{10^{14}}{1.5 \times 10^{10}}\right) = 0.228 \text{ V}$$

Then the work function difference is

$$\phi_{ms} = \phi'_m - \left(\chi' + \frac{E_g}{2e} + \phi_{fp}\right) = 3.20 - (3.25 + 0.555 + 0.228)$$

or

$$\phi_{ms} = -0.83 \text{ V}$$

■ Comment

The value of ϕ_{ms} will become more negative as the doping of the p-type substrate increases.

Degenerately doped polysilicon deposited on the oxide is also often used as the metal gate. Figure 11.13a shows the energy-band diagram of an MOS capacitor with an n$^+$ polysilicon gate and a p-type substrate. Figure 11.13b shows the energy-band diagram for the case of a p$^+$ polysilicon gate and the p-type silicon substrate. In the degenerately doped polysilicon, we will initially assume that $E_F = E_c$ for the n$^+$ case and $E_F = E_v$ for the p$^+$ case.

For the n$^+$ polysilicon gate, the metal-semiconductor work function difference can be written as

$$\boxed{\phi_{ms} = \left[\chi' - \left(\chi' + \frac{E_g}{2e} + \phi_{fp}\right)\right] = -\left(\frac{E_g}{2e} + \phi_{fp}\right)} \tag{11.12}$$

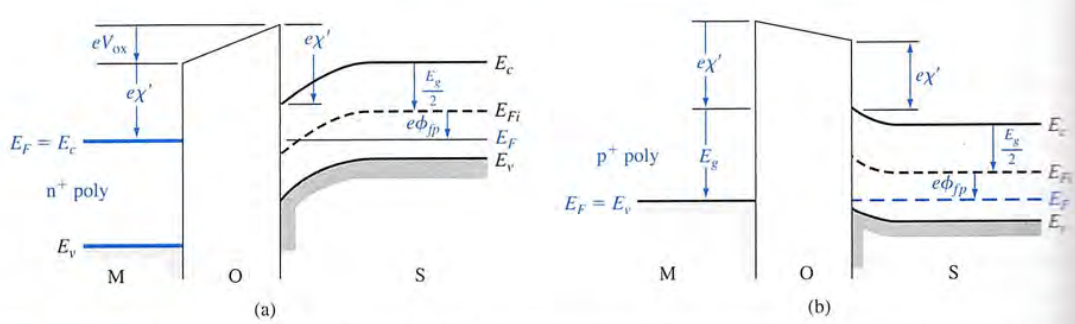

Figure 11.13 | Energy-band diagram through the MOS structure with a p-type substrate at zero gate bias for (a) an n$^+$ polysilicon gate and (b) a p$^+$ polysilicon gate.

and for the p^+ polysilicon gate, we have

$$\boxed{\phi_{ms} = \left[\left(\chi' + \frac{E_g}{e}\right) - \left(\chi' + \frac{E_g}{2e} + \phi_{fp}\right)\right] = \left(\frac{E_g}{2e} - \phi_{fp}\right)} \tag{11.13}$$

However, for degenerately doped n^+ polysilicon and p^+ polysilicon, the Fermi level can be above E_c and below E_v, respectively, by 0.1 to 0.2 V. The experimental ϕ_{ms} values will then be slightly different from the values calculated by using Equations (11.12) and (11.13).

We have been considering a p-type semiconductor substrate. We may also have an n-type semiconductor substrate in an MOS capacitor. Figure 11.14 shows the energy-band diagram of the MOS capacitor with a metal gate and the n-type semiconductor substrate, for the case when a negative voltage is applied to the gate. The metal–semiconductor work function difference for this case is defined as

$$\boxed{\phi_{ms} = \phi'_m - \left(\chi' + \frac{E_g}{2e} - \phi_{fn}\right)} \tag{11.14}$$

where ϕ_{fn} is assumed to be a positive value. We will have similar expressions for n^+ and p^+ polysilicon gates.

Figure 11.15 shows the work function differences as a function of semiconductor doping for the various types of gates. We may note that the magnitude of ϕ_{ms} for the polysilicon gates are somewhat larger than Equations (11.12) and (11.13) predict. This difference again is because the Fermi level is not equal to the conduction band

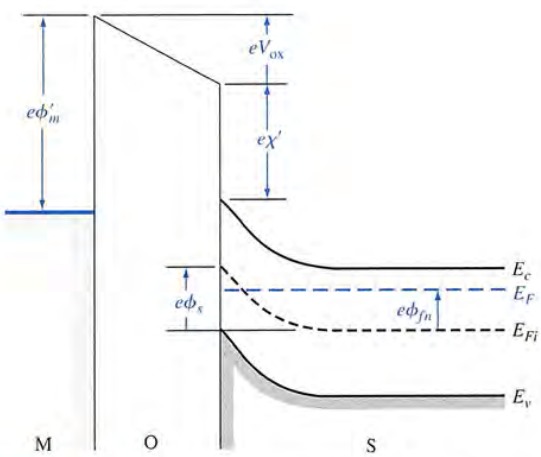

Figure 11.14 | Energy-band diagram through the MOS structure with an n-type substrate for a negative applied gate bias.

The equivalent oxide surface charge density is

$$Q'_{ss} = (10^{11})(1.6 \times 10^{-19}) = 1.6 \times 10^{-8}\ \text{C/cm}^2$$

The flat-band voltage is then calculated as

$$V_{FB} = \phi_{ms} - \frac{Q'_{ss}}{C_{ox}} = -1.1 - \left(\frac{1.6 \times 10^{-8}}{6.9 \times 10^{-8}}\right) = -1.33\ \text{V}$$

■ Comment

The applied gate voltage required to achieve the flat-band condition for this p-type substrate is negative. If the amount of fixed oxide charge increases, the flat-band voltage becomes even more negative.

TEST YOUR UNDERSTANDING

E11.6 Consider the MOS structure described in E11.3. For an oxide thickness of $t_{ox} =$ 200 Å and an oxide charge of $Q'_{ss} = 8 \times 10^{10}\ \text{cm}^{-2}$, calculate the flat-band voltage. (Ans. $V_{FB} = -1.90$ V)

E11.7 Repeat E11.6 for the MOS device described in E11.4. (Ans. $V_{FB} = -1.10$ V)

E11.8 Repeat E11.6 for the MOS device described in E11.5. (Ans. $V_{FB} = +0.105$ V)

11.1.5 Threshold Voltage

The threshold voltage was defined as the applied gate voltage required to achieve the threshold inversion point. The threshold inversion point, in turn, is defined as the condition when the surface potential is $\phi_s = 2\phi_{fp}$ for the p-type semiconductor and $\phi_s = 2\phi_{fn}$ for the n-type semiconductor. These conditions were shown in Figures 11.9a and 11.10. The threshold voltage will be derived in terms of the electrical and geometrical properties of the MOS capacitor.

Figure 11.18 shows the charge distribution through the MOS device at the threshold inversion point for a p-type semiconductor substrate. The space charge

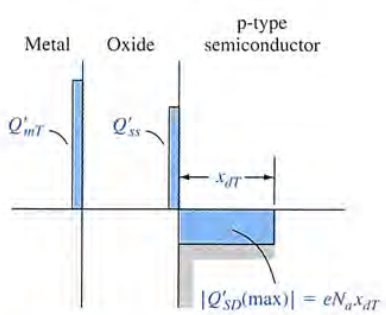

Figure 11.18 | Charge distribution in an MOS capacitor with a p-type substrate at the threshold inversion point.

width has reached its maximum value. We will assume that there is an equivalent oxide charge Q'_{ss} and the positive charge on the metal gate at threshold is Q'_{mT}. The prime on the charge terms indicates charge per unit area. Even though we are assuming that the surface has been inverted, we will neglect the inversion layer charge at this threshold inversion point. From conservation of charge, we can write

$$Q'_{mT} + Q'_{ss} = |Q'_{SD}(\text{max})| \tag{11.22}$$

where

$$\boxed{|Q'_{SD}(\text{max})| = eN_a x_{dT}} \tag{11.23}$$

and is the magnitude of the maximum space charge density per unit area of the depletion region.

The energy-band diagram of the MOS system with an applied positive gate voltage is shown in Figure 11.19. As we mentioned, an applied gate voltage will change the voltage across the oxide and will change the surface potential. We had from Equation (11.16) that

$$V_G = \Delta V_{\text{ox}} + \Delta\phi_s = V_{\text{ox}} + \phi_s + \phi_{ms}$$

At threshold, we can define $V_G = V_{TN}$, where V_{TN} is the threshold voltage that creates the electron inversion layer charge. The surface potential is $\phi_s = 2\phi_{fp}$ at threshold so Equation (11.16) can be written as

$$V_{TN} = V_{\text{ox}T} + 2\phi_{fp} + \phi_{ms} \tag{11.24}$$

where $V_{\text{ox}T}$ is the voltage across the oxide at this threshold inversion point.

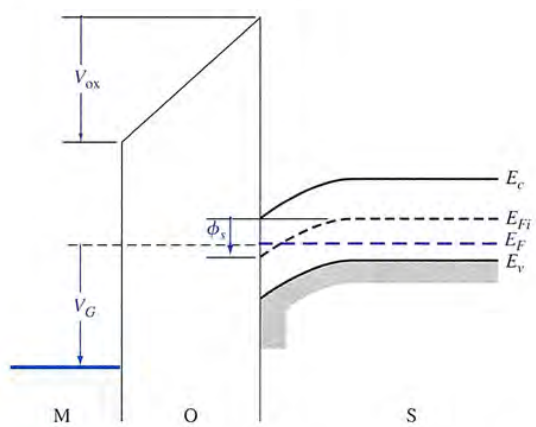

Figure 11.19 | Energy-band diagram through the MOS structure with a positive applied gate bias.

The voltage V_{oxT} can be related to the charge on the metal and to the oxide capacitance by

$$V_{oxT} = \frac{Q'_{mT}}{C_{ox}} \tag{11.25}$$

where again C_{ox} is the oxide capacitance per unit area. Using Equation (11.22), we can write

$$V_{oxT} = \frac{Q'_{mT}}{C_{ox}} = \frac{1}{C_{ox}}(|Q'_{SD}(\max)| - Q'_{ss}) \tag{11.26}$$

Finally, the threshold voltage can be written as

$$\boxed{V_{TN} = \frac{|Q'_{SD}(\max)|}{C_{ox}} - \frac{Q'_{ss}}{C_{ox}} + \phi_{ms} + 2\phi_{fp}} \tag{11.27a}$$

or

$$V_{TN} = (|Q'_{SD}(\max)| - Q'_{ss})\left(\frac{t_{ox}}{\epsilon_{ox}}\right) + \phi_{ms} + 2\phi_{fp} \tag{11.27b}$$

Using the definition of flat-band voltage from Equation (11.21), we can also express the threshold voltage as

$$V_{TN} = \frac{|Q'_{SD}(\max)|}{C_{ox}} + V_{FB} + 2\phi_{fp} \tag{11.27c}$$

For a given semiconductor material, oxide material, and gate metal, the threshold voltage is a function of semiconductor doping, oxide charge Q'_{ss}, and oxide thickness.

DESIGN EXAMPLE 11.4

Objective

To design the oxide thickness of an MOS system to yield a specified threshold voltage.

Consider an n^+ polysilicon gate and a p-type silicon substrate doped to $N_a = 3 \times 10^{16}$ cm^{-3}. Assume $Q'_{ss} = 10^{11}$ cm^{-2}. Determine the oxide thickness such that $V_{TN} = +0.65$ V.

■ Solution

From Figure 11.15, the work function difference is $\phi_{ms} = -1.13$ V. The various parameters can be calculated as

$$\phi_{fp} = V_t \ln\left(\frac{N_a}{n_i}\right) = (0.0259)\ln\left(\frac{3 \times 10^{16}}{1.5 \times 10^{10}}\right) = 0.376 \text{ V}$$

and

$$x_{dT} = \left\{\frac{4\epsilon_s \phi_{fp}}{eN_a}\right\}^{1/2} = \left\{\frac{4(11.7)(8.85 \times 10^{-14})(0.376)}{(1.6 \times 10^{-19})(3 \times 10^{16})}\right\}^{1/2} = 0.18\ \mu\text{m}$$

Then

$$|Q'_{SD}(\max)| = eN_a x_{dT} = (1.6 \times 10^{-19})(3 \times 10^{16})(0.18 \times 10^{-4})$$

or

$$|Q'_{SD}(\max)| = 8.64 \times 10^{-8} \text{ C/cm}^2$$

The oxide thickness can be determined from the threshold voltage equation

$$V_{TN} = (|Q'_{SD}(\max)| - Q'_{ss})\left(\frac{t_{ox}}{\epsilon_{ox}}\right) + \phi_{ms} + 2\phi_{fp}$$

Then

$$0.65 = \frac{[(8.64 \times 10^{-8}) - (10^{11})(1.6 \times 10^{-19})]}{(3.9)(8.85 \times 10^{-14})} \cdot t_{ox} - 1.13 + 2(0.376)$$

or

$$0.65 = 2.0 \times 10^5 t_{ox} - 0.378$$

which yields

$$t_{ox} = 504 \text{ Å}$$

■ Comment

The threshold voltage for this case is a positive quantity, which means that the MOS device is an enhancement mode device: a gate voltage must be applied to create the inversion layer charge, which is zero for zero applied gate voltage.

The threshold voltage must be within the voltage range of a circuit design. Although we have not yet considered the current in an MOS transistor, the threshold voltage is the point at which the transistor turns on. If a circuit is to operate between 0 and 5 V and the threshold voltage of a MOSFET is 10 V, for example, the device and circuit cannot be turned "on" and "off." The threshold voltage, then, is one of the important parameters of the MOSFET.

EXAMPLE 11.5

Objective

To calculate the threshold voltage of an MOS system using the aluminum gate.

Consider a p-type silicon substrate at $T = 300$ K doped to $N_a = 10^{14}$ cm^{-3}. Let $Q'_{ss} = 10^{10}$ cm^{-2}, $t_{ox} = 500$ Å, and assume the oxide is silicon dioxide. From Figure 11.15, we have that $\phi_{ms} = -0.83$ V.

■ Solution

We can start calculating the various parameters as

$$\phi_{fp} = V_t \ln\left(\frac{N_a}{n_i}\right) = (0.0259)\ln\left(\frac{10^{14}}{1.5 \times 10^{10}}\right) = 0.228 \text{ V}$$

and

$$x_{dT} = \left\{\frac{4\epsilon_s \phi_{fp}}{eN_a}\right\}^{1/2} = \left\{\frac{4(11.7)(8.85 \times 10^{-14})(0.228)}{(1.6 \times 10^{-19})(10^{14})}\right\}^{1/2} = 2.43 \ \mu\text{m}$$

that the semiconductor substrate is not inverted directly under the oxide with zero gate voltage. A positive gate voltage induces the electron inversion layer, which then "connects" the n-type source and the n-type drain regions. The source terminal is the source of carriers that flow through the channel to the drain terminal. For this n-channel device, electrons flow from the source to the drain so the conventional current will enter the drain and leave the source. The conventional circuit symbol for this n-channel enhancement mode device is also shown in this figure.

Figure 11.36 shows an n-channel depletion mode MOSFET. An n-channel region exists under the oxide with zero volts applied to the gate. However, we have shown that the threshold voltage of an MOS device with a p-type substrate may be

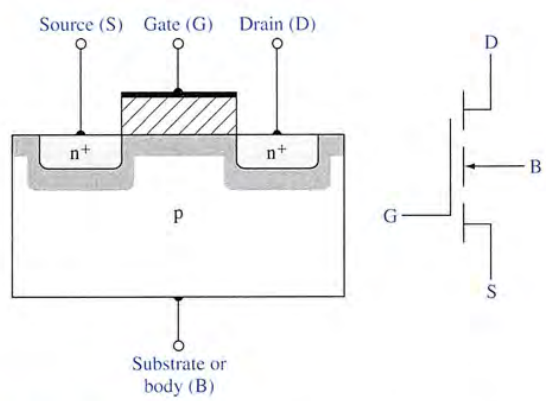

Figure 11.35 | Cross section and circuit symbol for an n-channel enhancement-mode MOSFET.

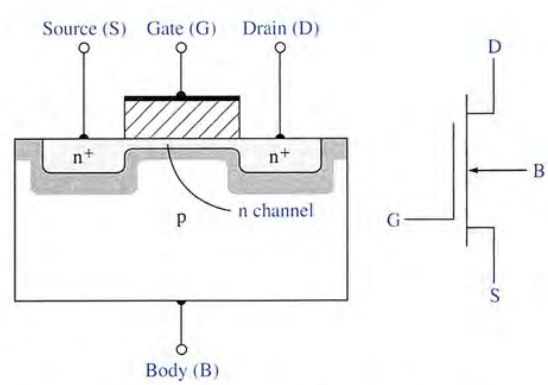

Figure 11.36 | Cross section and circuit symbol for an n-channel depletion-mode MOSFET.

negative; this means that an electron inversion layer already exists with zero gate voltage applied. Such a device is also considered to be a depletion mode device. The n-channel shown in this figure can be an electron inversion layer or an intentionally doped n-region. The conventional circuit symbol for the n-channel depletion mode MOSFET is also shown in the figure.

Figures 11.37a and 11.37b show a p-channel enhancement mode MOSFET and a p-channel depletion mode MOSFET. In the p-channel enhancement mode device, a negative gate voltage must be applied to create an inversion layer of holes that will "connect" the p-type source and drain regions. Holes flow from the source to the drain, so the conventional current will enter the source and leave the drain. A p-channel region exists in the depletion mode device even with zero gate voltage. The conventional circuit symbols are shown in the figure.

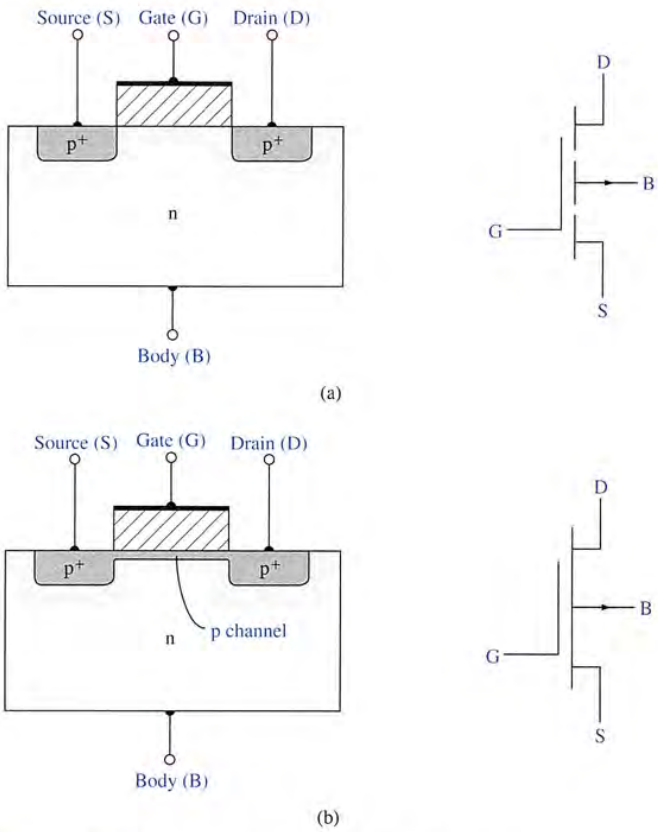

Figure 11.37 | Cross section and circuit symbol for (a) a p-channel enhancement mode MOSFET and (b) a p-channel depletion mode MOSFET.

11.3.2 Current–Voltage Relationship—Concepts

Figure 11.38a shows an n-channel enhancement mode MOSFET with a gate-to-source voltage that is less than the threshold voltage and with only a very small drain-to-source voltage. The source and substrate, or body, terminals are held at ground potential. With this bias configuration, there is no electron inversion layer, the drain-to-substrate pn junction is reverse biased, and the drain current is zero (disregarding pn junction leakage currents).

Figure 11.38b shows the same MOSFET with an applied gate voltage such that $V_{GS} > V_T$. An electron inversion layer has been created so that, when a small drain voltage is applied, the electrons in the inversion layer will flow from the source to the positive drain terminal. The conventional current enters the drain terminal and leaves the source terminal. In this ideal case, there is no current through the oxide to the gate terminal.

For small V_{DS} values, the channel region has the characteristics of a resistor, so we can write

$$I_D = g_d V_{DS} \tag{11.37}$$

where g_d is defined as the channel conductance in the limit as $V_{DS} \to 0$. The channel conductance is given by

$$g_d = \frac{W}{L} \cdot \mu_n |Q'_n| \tag{11.38}$$

where μ_n is the mobility of the electrons in the inversion layer and $|Q'_n|$ is the magnitude of the inversion layer charge per unit area. The inversion layer charge is a function of the gate voltage; thus, the basic MOS transistor action is the modulation of the channel conductance by the gate voltage. The channel conductance, in turn, determines the drain current. We will initially assume that the mobility is a constant; we will discuss mobility effects and variations in the next chapter.

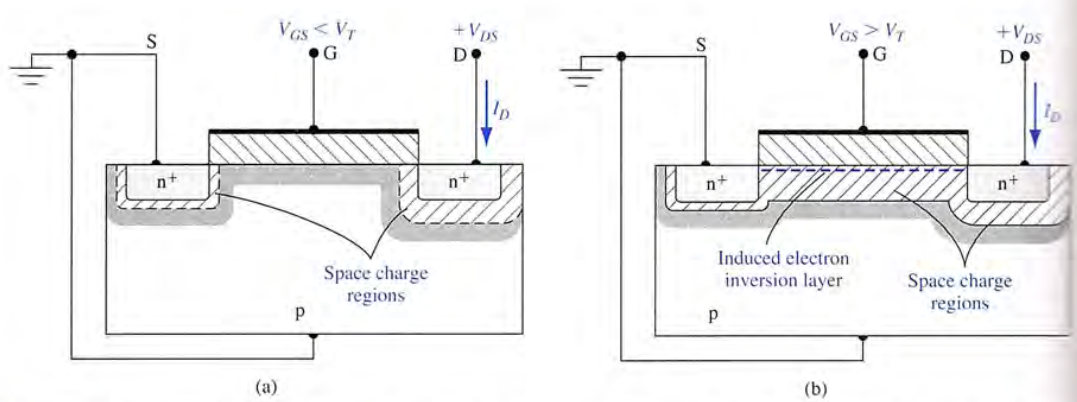

Figure 11.38 | The n-channel enhancement mode MOSFET (a) with an applied gate voltage $V_{GS} < V_T$, and (b) with an applied gate voltage $V_{GS} > V_T$.

The I_D versus V_{DS} characteristics, for small values of V_{DS}, are shown in Figure 11.39. When $V_{GS} < V_T$, the drain current is zero. As V_{GS} becomes larger than V_T, channel inversion charge density increases, which increases the channel conductance. A larger value of g_d produces a larger initial slope of the I_D versus V_{DS} characteristic as shown in the figure.

Figure 11.40a shows the basic MOS structure for the case when $V_{GS} > V_T$ and the applied V_{DS} voltage is small. The thickness of the inversion channel layer in the figure qualitatively indicates the relative charge density, which is essentially constant along the entire channel length for this case. The corresponding I_D versus V_{DS} curve is shown in the figure.

Figure 11.40b shows the situation when the V_{DS} value increases. As the drain voltage increases, the voltage drop across the oxide near the drain terminal decreases, which means that the induced inversion charge density near the drain also decreases. The incremental conductance of the channel at the drain decreases, which then means that the slope of the I_D versus V_{DS} curve will decrease. This effect is shown in the I_D versus V_{DS} curve in the figure.

When V_{DS} increases to the point where the potential drop across the oxide at the drain terminal is equal to V_T, the induced inversion charge density is zero at the drain terminal. This effect is schematically shown in Figure 11.40c. At this point, the incremental conductance at the drain is zero, which means that the slope of the I_D versus V_{DS} curve is zero. We can write

$$V_{GS} - V_{DS}(\text{sat}) = V_T \tag{11.39a}$$

or

$$\boxed{V_{DS}(\text{sat}) = V_{GS} - V_T} \tag{11.39b}$$

where $V_{DS}(\text{sat})$ is the drain-to-source voltage producing zero inversion charge density at the drain terminal.

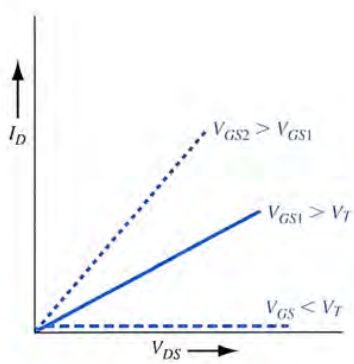

Figure 11.39 | I_D versus V_{DS} characteristics for small values of V_{DS} at three V_{GS} voltages.

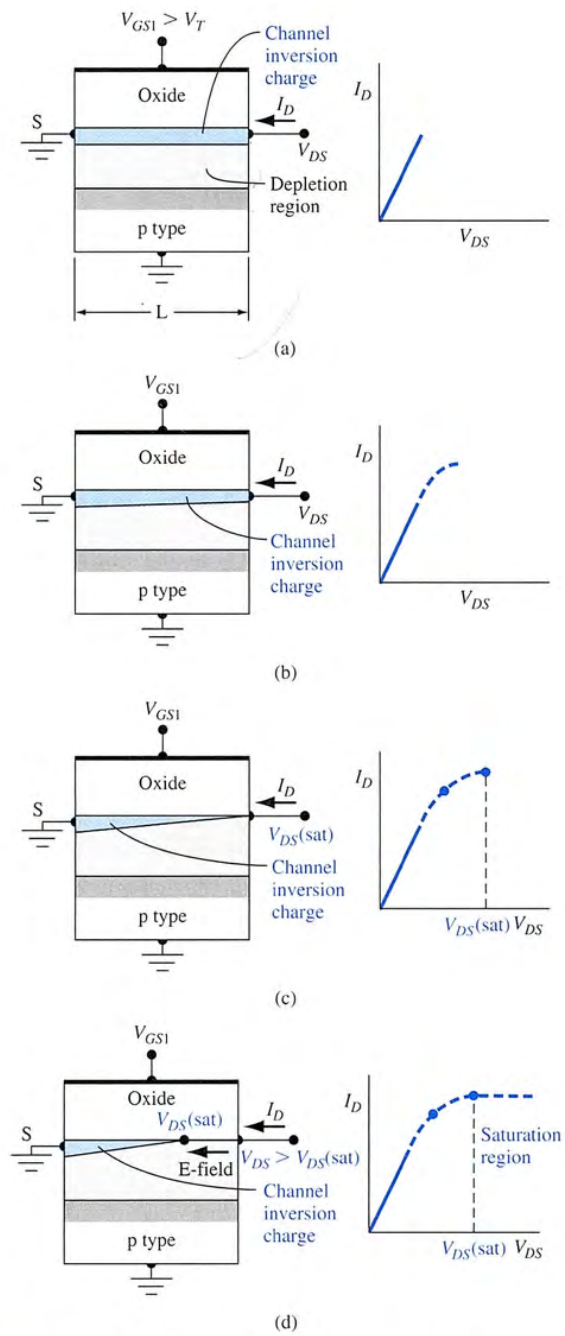

Figure 11.40 | Cross section and I_D versus V_{DS} curve when $V_{GS} < V_T$ for (a) a small V_{DS} value, (b) a larger V_{DS} value, (c) a value of $V_{DS} = V_{DS}$(sat), and (d) a value of $V_{DS} > V_{DS}$(sat).

When V_{DS} becomes larger than the $V_{DS}(\text{sat})$ value, the point in the channel at which the inversion charge is just zero moves toward the source terminal. In this case, electrons enter the channel at the source, travel through the channel toward the drain, and then, at the point where the charge goes to zero, the electrons are injected into the space charge region where they are swept by the E-field to the drain contact. If we assume that the change in channel length ΔL is small compared to the original length L, then the drain current will be a constant for $V_{DS} > V_{DS}(\text{sat})$. The region of the I_D versus V_{DS} characteristic is referred to as the *saturation region.* Figure 11.40d shows this region of operation.

When V_{GS} changes, the I_D versus V_{DS} curve will change. We saw that, if V_{GS} increases, the initial slope of I_D versus V_{DS} increases. We can also note from Equation (11.39b) that the value of $V_{DS}(\text{sat})$ is a function of V_{GS}. We can generate the family of curves for this n-channel enhancement mode MOSFET as shown in Figure 11.41.

Figure 11.42 shows an n-channel depletion mode MOSFET. If the n-channel region is actually an induced electron inversion layer created by the metal–semiconductor work function difference and fixed charge in the oxide, the current–voltage characteristics are exactly the same as we have discussed, except that V_T is a negative quantity. We may also consider the case when the n-channel region is actually an n-type semiconductor region. In this type of device, a negative gate voltage will induce a space charge region under the oxide, reducing the thickness of the n-channel region. The reduced thickness decreases the channel conductance, which reduces the drain current. A positive gate voltage will create an electron accumulation layer, which increases the drain current. One basic requirement for this device is that the channel thickness t_c must be less than the maximum induced space charge width in order to be able to turn the device off. The general I_D versus V_{DS} family of curves for an n-channel depletion mode MOSFET is shown in Figure 11.43.

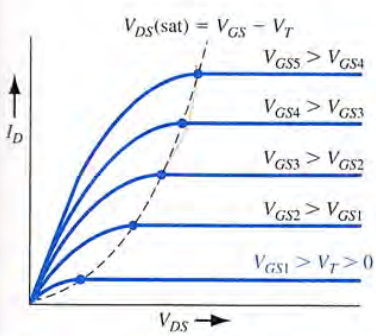

Figure 11.41 | Family of I_D versus V_{DS} curves for an n-channel enhancement-mode MOSFET.

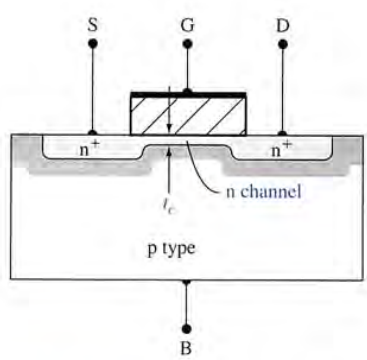

Figure 11.42 | Cross section of an n-channel depletion-mode MOSFET.

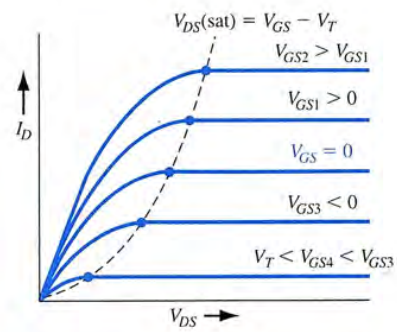

Figure 11.43 | Family of I_D versus V_{DS} curves for an n-channel depletion-mode MOSFET.

In the next section we will derive the ideal current–voltage relation for the n-channel MOSFET. In the nonsaturation region, we will obtain

$$\boxed{I_D = \frac{W\mu_n C_{ox}}{2L}\left[2(V_{GS} - V_T)V_{DS} - V_{DS}^2\right]} \tag{11.40}$$

and, in the saturation region, we will have

$$\boxed{I_D = \frac{W\mu_n C_{ox}}{2L}(V_{GS} - V_T)^2} \tag{11.41}$$

The operation of a p-channel device is the same as that of the n-channel device, except the charge carrier is the hole and the conventional current direction and voltage polarities are reversed.

*11.3.3 Current–Voltage Relationship—Mathematical Derivation

In the previous section, we qualitatively discussed the current–voltage characteristics. In this section, we will derive the mathematical relation between the drain current, the gate-to-source voltage, and the drain-to-source voltage. Figure 11.44 shows the geometry of the device that we will use in this derivation.

In this analysis, we will make the following assumptions:

1. The current in the channel is due to drift rather than diffusion.
2. There is no current through the gate oxide.
3. A gradual channel approximation is used in which $\partial E_y/\partial y \gg \partial E_x/\partial x$. This approximation means that E_x is essentially a constant.
4. Any fixed oxide charge is an equivalent charge density at the oxide-semiconductor interface.
5. The carrier mobility in the channel is constant.